Iratxe Bermozabal Campillo.
10.10.96.

CAPÍTULO II

DISTRIBUCIONES MUESTRALES DE PROBABILIDAD

2.1. Distribuciones muestrales

2.2. Distribución muestral de medias

2.3. El teorema central del límite

2.4. Distribución de la varianza muestral (varianza en todas las muestras de tamaño 'n').

2.5. Distribución muestral de la diferencia de medias

2.6. Distribución muestral de la proporción de éxitos

2.7. Distribución muestral de la diferencia de proporciones

TEXTO: Apuntes Fotocopiados

BIBLIOGRAFÍA COMPLEMENTARIA:

BERENSON, M. L. y LEVINE, D. L.: *Estadística para administración y economía*, Madrid, Interamericana, 1987; capítulo 7, págs. 205-227.

CANAVOS, G. C.: *Probabilidad y estadística*, Madrid, McGraw-Hill, 1988; capítulo 7, págs.214-247.

MILLS, R. L.: *Estadística para economía y administración*, Madrid, McGraw-Hill, 1980; capítulo 5, págs.137-173.

SPIEGEL, M. R.: *Estadística*, Madrid, McGraw-Hill, 1990; capítulo 8, págs.186-207.

DISTRIBUCIONES MUESTRALES DE PROBABILIDAD: EJEMPLOS

EJEMPLO 1:

Supongamos una variable aleatoria X con posibles valores 1, 2, 3 y 4. Elegimos al azar una muestra de tamaño 2. ¿Cuántas muestras posibles podemos obtener? ¿Qué valores posibles puede tener el estadístico $\overline{X}$? ¿Cuáles serán sus probabilidades?

EJEMPLO 2:

Una compañía telefónica de cierta región de Estados Unidos, ha solicitado un incremento de tarifas de todos los teléfonos residenciales, incluyendo un 25% de incremento en los teléfonos de los estudiantes. Para oponerse a este aumento, un grupo de estudiantes decide investigar los costos telefónicos en la ciudad. De acuerdo con la información proporcionada por la compañía, el pago mensual promedio de todos los abonados de la ciudad ha sido de 15,30 $, con una desviación estándar de 4,10 $, y los estudiantes quieren saber si la situación es análoga para sus teléfonos. Puesto que no disponen de la información necesaria, deciden tomar una muestra aleatoria de n = 36. Si en la muestra se obtiene un pago promedio de 14,00 $, ¿qué podrías decir al respecto?

EJEMPLO 3:

Obtener los puntos críticos de las correspondientes:

a) X^2 0,95; 16.

b) X^2 0,80; 10.

EJEMPLO 4:

Hallar la probabilidad de que una variable aleatoria X que sigue una distribución X^2 con 11 grados de libertad, sea mayor que 20.

EJEMPLO 5:

Hallar la probabilidad de que una variable aleatoria X que sigue una distribución X^2 con 10 grados de libertad, se encuentre comprendida entre los valores 4 y 6.

EJEMPLO 6:

Obtener el punto crítico para un nivel de significación $\alpha = 0,05$ de una variable aleatoria X que sigue una distribución t de Student con 10 grados de libertad.

EJEMPLO 7:

Obtener el punto crítico para un nivel de significación $\alpha = 0,30$ de una variable aleatoria X que sigue una distribución t de Student con 12 grados de libertad.

EJEMPLO 8:

Lo mismo que en el **Ejemplo 7**, para un nivel de significación $\alpha = 0,80$.

EJEMPLO 9:

Hallar la probabilidad de que una variable aleatoria X que sigue una distribución t de Student con 10 grados de libertad sea mayor que 2.

EJEMPLO 10:

Hallar la probabilidad de que una variable aleatoria X que sigue una distribución t de Student con 5 grados de libertad esté comprendido entre 1 y 3.

EJEMPLO 11:

Una pequeña compañía financiera que ha informado a una firma auditora que sus préstamos pendientes de pago se distribuyen aproximadamente en forma normal con una media de 825 $, desconociéndose su desviación estándar. Para verificar lo dicho acerca de la media, se toma una muestra aleatoria de 25 cuentas de clientes, resultando que $\overline{X} = 780$ $ y $S = 105$ $. ¿Qué podría decirse respecto al dato facilitado por la compañía financiera?

EJEMPLO 12:

Los diámetros en pulgadas de ciertas bolas de acero para cojinetes, deben tener una varianza $0,0010 = \sigma_X^2$ pulgadas^2. Si se toma una muestra aleatoria de n = 21,

queremos comprobar hasta qué punto la afirmación de que la varianza sea de 0,0010 es cierta si obtenemos que la varianza muestral es de 0,0016.

EJEMPLO 13:

Calcular el punto crítico para una variable aleatoria X que sigue una distribución F 0,05 (12; 18).

EJEMPLO 14:

Calcular la probabilidad de que una variable aleatoria X que sigue una distribución F de Fishe-Snedecor con 12 y 15 grados de libertad, esté comprendida entre 2,5 y 3.

PROBLEMAS DE ESTADÍSTICA

1º) Una población se compone de los cinco números 2, 3, 6, 8, 11. Considerar todas las muestras posibles de tamaño 2 que pueden extraerse con reemplazamiento de esta población. Hallar:

a) La media de la población.

b) La desviación típica de la población. $\sigma^2 = E(x^2) - (E(x))^2$

c) La media de la distribución muestral de medias.

d) La desviación típica de la distribución muestral de medias. $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$

2º) Ciertas piezas deben ser enviadas en cajas, dentro de las cuales quedan bien ajustadas. La longitud de las piezas, expresadas en cm., se distribuye normalmente $N(\mu, \sigma^2) = N(20; 0,04)$ y la de las cajas se distribuye normalmente $N(\mu, \sigma^2) = N(20, 615; 0,05)$. Elegidas al azar una pieza y una caja, ¿cuál es la probabilidad de que:

a) La pieza no quepa en la caja?

b) La longitud de la caja supere a la de la pieza en 0,231 cm. por lo menos?

3º) Lo mismo que en el **problema 1º)** para el caso de muestreo sin reemplazamiento.

4º) El control de calidad de cierto tipo de piezas es destructivo. El coste en pesetas de cada pieza destruida al ser sometida al correspondiente control es 100 veces

el valor de su longitud en mm. Ésta última se distribuye normalmente con media 12 y varianza 0,04. ¿Cómo se distribuirá la pérdida media por pieza inspeccionada (y destruida) a partir de muestras aleatorias de tamaño 16? ¿Cuál es la probabilidad de que dicha pérdida media diaria sea mayor que 1.203 pesetas?

5º) La variable aleatoria X tiene media μ y varianza 100. Se toma una muestra aleatoria de X que nos da una $\overline{X}$. Hallar el tamaño de la muestra n que se necesita para que la probabilidad de $\overline{X}$ sea por lo menos del 0,75 de que se encuentre a menos de 0,1 de μ.

6º) Sea X una variable aleatoria cuyos valores equiprobables son 0, 1, 4, 6. Eligiendo muestras aleatorias simples binarias (con reposición), calcular la media y la varianza de $\overline{X}$.

7º) La duración de las llamadas telefónicas a larga distancia sigue una distribución normal con μ = 8 minutos, σ = 2 minutos. Si se seleccionan muestras aleatorias de 25 llamadas:

a) Calcular el error estándar de la distribución muestral de medias.

b) ¿Qué porcentaje de medias muestrales estaría entre 7,8 y 8,2 minutos?

Si se seleccionan muestras aleatorias de 100 llamadas:

c) ¿Qué porcentaje de medias muestrales estaría entre 7,8 y 8,2 minutos?

d) Explicar la diferencia entre los resultados de **b)** y **d)**.

8º) Una máquina automática fabrica bolas de billar, de las que se quiere controlar su diámetro X. La cantidad X es una variable aleatoria, cuyos parámetros son $\mu = 20$ mm. y $\sigma = 1$ mm. Se extraen, a intervalos regulares, muestras de tamaño $n = 100$; los diámetros X de las bolas extraídas se miden. Sea $\overline{X}$ la media de los diámetros de una muestra dada:

a) ¿Cuál es la naturaleza de la extracción efectuada?

b) ¿Qué distribución sigue la variable aleatoria $\overline{X}$?

c) Determinar $E(\overline{X})$ y $\sigma(\overline{X})$.

9º) Las cuentas de gastos de comidas de los ejecutivos de una agencia de publicidad, tienen una media de 10 $ por persona y una desviación estándar de 4 $ por persona. Si se seleccionan muestras aleatorias de 16 cuentas:

a) ¿Por debajo de qué valor en dinero caerá el 99% de las medias muestrales?

b) ¿Qué porcentaje de medias muestrales estará entre 8 $ y 12 $?

d) ¿Qué suposición se debe hacer a fin de resolver **a)** y **b)**?

Si se seleccionan muestras aleatorias de 64 cuentas:

e) ¿Hay que hacer alguna suposición para resolver **d)**?

f) Comparar los resultados obtenidos en **a)** y **d)**.

10º) En la producción de cierto material se sabe que la desviación estándar de la tensión de ruptura de este material es de 25 libras. ¿Cuál debe ser la tensión de ruptura promedio del proceso si, con base en una muestra aleatoria de 50 especímenes, la probabilidad de que la media muestral tenga un valor mayor de 250 libras es 0,95?

11º) Dados dos conjuntos A y B, $A = \{1,3,4\}$ y $B = \{2,6\}$:

a) Formar la distribución muestral de las diferencias $D = (A - B)$.

b) Calcular la media de D.

c) Calcular la desviación típica de D.

12º) Para cierta prueba de aptitud se sabe, con base en la experiencia, que el número promedio de aciertos es 1.000 con una desviación típica de 125. Si se aplica la prueba a 100 personas seleccionadas al azar, calcular las siguientes probabilidades:

a) $P(985 < \overline{X} < 1.015)$.

b) $P(\overline{X} > 1.020)$.

c) $P(\overline{X} < 975)$.

13º) Siendo $\overline{X}$ la media de una muestra aleatoria simple de tamaño 9, extraída de una población con distribución normal de media 0 y varianza 81: determinar K tal que $P(\overline{X} > k) = 0{,}20$.

14º) Un contratista piensa comprar una gran cantidad de lámparas de alta intensidad a cierto fabricante. Éste asegura que la duración promedio de las lámparas es de 1.000 horas, con una desviación estándar igual a 80 horas. El contratista decide comprar las lámparas sólo si una muestra aleatoria de 64 de éstas da como resultado una vida promedio de por lo menos 1.050 horas. ¿Cuál es la probabilidad de que el contratista adquiera las lámparas?

15º) Una máquina empaqueta un cierto producto en dosis de peso X, donde X está distribuido normalmente con media de 25 grs. y desviación típica de 0,4 grs.; el peso del paquete vacío Y sigue también una distribución normal con media de 5 grs. y desviación típica 0,2 grs. Si se toma al azar una muestra de 100 paquetes, ¿cuál es la probabilidad de que el peso medio de los 100 paquetes esté comprendido entre 89 y 94 grs.?

16º) Las bombillas eléctricas de un fabricante A tienen una duración media de 1.400 horas con una desviación estándar de 200 horas, mientras que las de otro fabricante B tienen una duración media de 1.200 horas

con una desviación estándar de 100 horas. Si se toman muestras al azar de 125 bombillas de cada fabricante, ¿cuál es la probabilidad de que las bombillas de A tengan una duración media que sea al menos (a) 160 horas /(b) 250 horas más que las bombillas de B?

17º) Un inspector federal de pesos y medidas visita una planta de empacado para verificar que el peso neto de las cajas sea el indicado en éstas. El gerente de la planta asegura al inspector que el peso promedio de cada caja es de 750 grs., con una desviación estándar de 5 grs. El inspector selecciona al azar 100 cajas y encuentra que el peso promedio es de 748 grs. Bajo estas condiciones, ¿qué tan probable es tener un peso promedio de 748 grs. o menos?; ¿qué actitud debe tomar el inspector?

18º) En la fabricación de cojinetes para motores, se sabe que el diámetro promedio es de 5 cm. con una desviación estándar igual a 0,005 cm. El proceso es vigilado de forma periódica mediante la selección aleatoria de 64 cojinetes, midiendo sus correspondientes diámetros. El proceso no se detiene mientras la probabilidad de que la media muestral se encuentre entre dos límites especificados sea de 0,95. Determinar el valor de estos límites.

19º) Una población E está formada por los elementos $E = \{1,2,4,6\}$:

a) Calcular la proporción p de las cifras impares.

b) Escribir todas la muestras de tamaño 2 que puedan extraerse, de forma no exhaustiva, de la población E.

c) Calcular para cada una de las muestras anteriores la proporción f de las cifras impares.

d) Calcular la media y la desviación típica de la distribución muestral de las proporciones f.

20º) Si se obtiene una muestra aleatoria de n = 16 de una distribución normal con media y varianza desconocidas, obtener: $P(S^2 / \sigma^2 < 2,041)$.

SOLUCIONES

1º) a) 6; **b)** 3,2863; **c)** 6; **d)** 2,32. **2º) a)** 0,0203; **b)** 0,1003.
3º) 6; 2,012. **4º)** N(1.200; 5); 0,2743. **5º)** n = 40.000.
6º) 11/4; 2,84. **7º) a)** 0,4; **b)** 0,383; **c)** 0,7026.
8º) a) m. a. s. con n = 100; **b)** Normal; **c)** 20; $1/\sqrt{100}$.
9º) a) 12,33 $; **b)** 0,9544; **d)** 11.165 $. **10º)** 255,82.
11º) b) - 4/3; **c)** 2,3570.
12º) a) 0,7698; **b)** 0,0548; **c)** 0,0228. **13º)** k = 2,52.
14º) 0. **15º)** 0. **16º) a)** 0,6517; **b)** 0,3121. **17º)** 0.
18º) (4,998; 5,001). **19º) a)** 1/4; **d)** 1/4; 1/4. **20º)** 0,99.

EJEMPLO DE DISTRIBUCIÓN MUESTRAL

EJEMPLO 1.

* Posibles valores de la variable aleatoria X: 1; 2; 3; 4.
* Elegimos una muestra de n = 2.

MUESTRA	X_1	X_2	$\overline{X}$
1	1	1	1
2	1	2	1,5
3	1	3	2
4	1	4	2,5
5	2	2	2
6	2	1	1,5
7	2	3	2,5
8	2	4	3
9	3	3	3
10	3	1	2
11	3	2	2,5
12	3	4	3,5
13	4	4	4
14	4	1	2,5
15	4	2	3
16	4	3	3,5

$\overline{X}$ calculo la media en cada uno de estas muestras.

Son ≠ debido a la importancia del orden. Muy importante en términos de probabilidad.

↳ sólo hasta 16 pq. me planteo el nº de muestras ≠ que puedo obtener cuando la población tiene 4 elementos, (importa el orden).

VALORES DIFERENTES DE $\overline{X}$ (MEDIA MUESTRAL)	Nº DE MUESTRAS	PROBABILIDADES
1	1	1/16
1,5	2	2/16
2	3	3/16
2,5	4	4/16
3	3	3/16
3,5	2	2/16
4	1	1/16

* Resumiendo

N (POBLACIÓN)

↓

EXTRAEMOS MUESTRAS DE TAMAÑO **n**

muestras	n_1	n_2	.....	n_m
elementos	$x_1 \dots x_n$	$x_1 \dots x_n$		$x_1 \dots x_n$
	↓	↓		↓
obtenemos las medias muestrales:	$\overline{X}_1$	$\overline{X}_2$		$\overline{X}_m$

Un estadístico es cualquier función de una variable aleatoria. Esta función no va a poseer cantidades desconocidas.

Capítulo 2

MEDIA DE LA DISTRIBUCIÓN MUESTRAL DE MEDIAS

→ media de las medias muestrales por medio de la esperanza matemática.

La distribución muestral (media muestral) es un modelo teórico. nos expresa una relación entre cada uno del estadístico muestral y el correspon. $\mu_{\bar{X}} = \Sigma\ \bar{X}_i\ P(\bar{X}_i)$ parámetro poblacional.

$$\mu_{\bar{X}} = \overset{\bar{x}_i \cdot P(\bar{x}_i)}{1 \cdot 1/16 + 1{,}5 \cdot 2/16 + \ldots + 4 \cdot 1/16} = 2{,}5$$

por ser igual el valor podemos decir que el estimador es insesgado

PARA LA POBLACIÓN:

$$E(X) = \overset{x_i \quad P(x_i)}{1 \cdot 1/4 + 2 \cdot 1/4 + 3 \cdot 1/4 + 4 \cdot 1/4} = 2{,}5$$

VARIANZA DE LA DISTRIBUCIÓN MUESTRAL DE MEDIAS

$\bar{X}$ (→ media muestral)	$P(\bar{X})$	$(\bar{X} - \mu_{\bar{X}})$ (2,5)	$(\bar{X} - \mu_{\bar{X}})^2$	$(\bar{X} - \mu_{\bar{X}})^2 \cdot P(\bar{X})$
1	1/16	-1,5	2,25	2,25/16 → varianza de la media muestral
1,5	2/16	-1	1	2/16
2	3/16	-0,5	0,25	0,75/16
2,5	4/16	0	0	0
3	3/16	0,5	0,25	0,75/16
3,5	2/16	1	1	2/16
4	1/16	1,5	2,25	2,25/16
				0,625

* La varianza de la media muestral siempre será ≤ que la varianza poblacional.
* La varianza de la media muestral es 1 / nveces la varianza poblacional.

→ por ser una v. aleatoria, tendrá su ley de probabilidad.

$$\sigma^2_{\bar{X}} = E\left[(\bar{X} - \mu_{\bar{X}})^2\right] \qquad \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

→ media obtenida en la muestra. (del cual se pueden obtener ≠ valores y/o diferentes probabilidades).

La media muestral es un estadístico utilizado para hacer inferencias sobre el parámetro poblacional (= la verdadera media de la población).

Hay que plantearse si el estimador $\bar{X}$ es bueno o no, sino por qué no obtengo otro estimador (ej. → mediana).

RESUMEN

Si X → Normal ➔ $\bar{X}$ → Normal ($\forall n$)

Si X ↛ Normal ➔ $\bar{X}$ → Normal ($\forall n \geq 30$) **TCL** (Teorema central del límite)

Si el tamaño de la muestreo es superior al 10% de la población debemos aplicar el factor corrector para poblaciones finitas.

muestreo con reposición

POBLACIÓN	MCR	MSR
FINITA (la dispersión ↓)	$\bar{X} \rightarrow N(\mu_x; \frac{\sigma_x}{\sqrt{n}})$ (media; desviación típica)	$\bar{X} \rightarrow N(\mu_x; \frac{\sigma_x}{\sqrt{n}} \underbrace{\sqrt{\frac{N-n}{N-1}}}_{\text{factor corrector}})$
INFINITA	$\bar{X} \rightarrow N(\mu_x; \frac{\sigma_x}{\sqrt{n}})$	$\bar{X} \rightarrow N(\mu_x; \frac{\sigma_x}{\sqrt{n}})$

¡OJO! * Teorema central del límite (en el libro): según vayamos ↑ el tamaño de la muestra y para un tamaño suficiente/grande ≥ 30, la distribución media muestral seguirá una distribución normal.

*σ conocida

4 PROPIEDADES DE UN BUEN ESTIMADOR

1. Ser INSESGADO (cuando su valor esperado sea igual al verdadero valor del parámetro que se está estimando. E[estimador] = Parámetro poblacional).
2. Ser EFICIENTE (es 1 prop. relativa. 1º estimador es más eficiente que otro cuando tiene menor varianza).
3. Ser SUFICIENTE (cuando incluye toda la información que la muestra pueda proporcionar a cerca del parámetro poblacional).
4. Ser CONSISTENTE (cuando proporciona estimaciones que *convergen en probabilidad hacia el parámetro poblacional).

* CONVERGENCIA → Ejem.: tirar 1 moneda al aire: 50 veces: 24c; 26+ / 100 " : 47c; 53+ / 1000 " : 506c; 494+ / 1.000.000 ".

↗ cuanto > es 'n' esta proporción entre caras y cruces será el 50%. pero no se garantiza que esa prop se dé continua/a partir de un valor

Capítulo 2

* **Población:** v.a. X → B(n; p)

* **m.a.s.:** De N → n

* **Podemos generar:**

  1. Distribución muestral del "Nº de Éxitos"
  2. Distribución muestral de la "Proporción de Éxitos"

* **Vamos a definir:**

  R → Nº de éxitos en la población

  r → Nº de éxitos en la muestra

$$\Pi = R/N$$

  $\overline{P}$ → Proporción de éxitos en la muestra

$$\overline{P} = r/n$$

Chi-cuadrado.

* Distribución χ^2_n para una v.a. X derivada de la distribución muestral (pág. 24)

"Es una suma de los cuadrados de 'n' variables aleatorias, todas ellas con distribuciones normales(0,1)." (n sumandos: n grados de libertad => se refiere al nº de valores q pueden ser arbitrarios/escogidos).

$\chi^2_n = Z_1^2 + Z_2^2 + \dots + Z_n^2$. $[Z \sim N(0,1)]$

DISTRIBUCIÓN MUESTRAL DEL ''Nº DE ÉXITOS''

* **Población:** v.a. $X \sim B(n; \pi)$

$\downarrow$

m.a.s.: n

* **v.a.** r $\rightarrow$ ''Nº de éxitos en una muestra de tamaño n''

* **¿Posibles valores?:** $0 \leq r \leq n$

* **¿Probabilidades?:** $P[r=k] = \dfrac{n!}{k!\,(n-k)!}\ (\pi)^k\ (1-\pi)^{n-k}$

* **E[r]** $= \mu_r = n\ \pi$

* **Var(r)** $= \sigma_r^2 = n\ \pi\ (1-\pi)$

DISTRIBUCIÓN MUESTRAL DE LA PROPORCIÓN

Distribución muestral se limita al muestreo aleatorio simple. Hay riesgo de posibles errores.

* **Población:** v.a. $X \sim B(n; \pi)$

↓

m.a.s.: n

* **¿Posibles valores?**

v.a. $\overline{P}$ → "Proporción de éxitos"

(n + 1) Posibles valores

0/n; 1/n; ....; (n-1)/n; n/n

* **¿Probabilidades?**

$$P[\overline{p}=k] = \frac{n!}{k!\,(n-k)!}\,(\pi)^k\,(1-\pi)^{n-k}$$

* $E[\overline{P}] = \mu_{\overline{P}} = \pi$

* $Var[\overline{P}] = \sigma^2_{\overline{P}} = \frac{\pi\,(1-\pi)}{n}$

* **Aproximación mediante distribución normal:**

1. n grande
2. $n\,\pi\,(1-\pi) \geq 3$

* **Corrección por continuidad:**

$\overline{P} \pm 1/2n$

(Tablas: pág. 170)

RESUMEN DE CÁLCULOS PARA MUESTRAS DE TAMAÑO 2

			Primera Selección		
	r		0	1	←Número de personas que usan lentes de contacto en la primera selección
		P(r)	0.7	0.3	
Segunda Selección	0	0.7	0 [0.00] 0.49	1 [0.50] 0.21	←Número de usuarios en muestra de tamaño 2 : r ←Proporción de la muestra; $\overline{p}$ = r / n ←Probabilidad de ocurrir de esta muestra
	1	0.3	1 [0.50] 0.21	2 [1.00] 0.09	

DISTRIBUCIONES MUESTRALES PARA r Y $\overline{p}$, n = 2

Variable aleatoria		Probabilidad
Número de personas que usan lentes de contacto (r)	Proporción de personas que usan lentes de contacto ($\overline{p} = r / n$)	P (r) o P ($\overline{p}$)
0	0.00	0.49
1	0.50	0.42
2	1.00	0.09
		1.00

CÁLCULO DE LA MEDIA Y DE LA VARIANZA PARA CADA DISTRIBUCIÓN MUESTRAL

Número de éxitos			Proporción de éxitos		
$r \cdot P(r)$	$(r-\mu_r)^2$	$P(r)\cdot(r-\mu_r)^2$	$\bar{p}\cdot P(\bar{p})$	$(\bar{p}-\mu_{\bar{p}})^2$	$P(\bar{p})\cdot(\bar{p}-\mu_{\bar{p}})^2$
0.00	0.36	0.1764	0.00	0.09	0.0441
0.42	0.16	0.0672	0.21	0.04	0.0168
0.18	1.96	0.1764	0.09	0.49	0.0441
0.60		0.4200	0.30		0.1050
$= E(r)$		$= Var(r)$	$= E(\bar{p})$		$= Var(\bar{p})$
$= \mu_r$			$= \mu_{\bar{p}}$		

Distribución de la razón de Varianzas Muestrales

Condiciones

* Población: v.a. $x \rightarrow N(\mu_x; \sigma_x^2)$ independiente.
  v.a. $y \rightarrow N(\mu_y; \sigma_y^2)$

* Muestra: Eligimos dos muestras:
  $n_x \rightarrow$ m.a.s. de la población x.
  $n_y \rightarrow$ m.a.s. de la población y.

* $\dfrac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \rightarrow F_{n_x-1;\,n_y-1}$.

Si $\sigma_x^2 = \sigma_y^2$ la probabilidad de verificar 1 valor $f \neq 1$ es pequeña.

Prueba de Homocedasticidad.

Si $\sigma_x^2 = \sigma_y^2$, trabajo con el cociente: S_x^2/S_y^2.

* Si el valor de este cociente es de manera aproximada $\neq 1 \rightarrow \sigma_x^2 \neq \sigma_y^2$

Ej.: $\dfrac{S_x^2}{S_y^2} = 2{,}13$ Pto crítico de $F_{n_x-1;\,n_y-1}$

$P[F > 2{,}13]$

RESUMEN DE CÁLCULOS PARA MUESTRAS DE TAMAÑO 3

			Primeras Dos Selecciones		
	r		0	1	2
		P(r)	0.49	0.42	0.09
Tercera Selección	0	0.7	0 [0.00] 0.343	1 [0.33] 0.294	2 [0.67] 0.063
	1	0.3	1 [0.33] 0.147	2 [0.67] 0.126	3 [1.00] 0.027

←Número de éxitos, r
←Proporción de éxitos, r / n
←Probabilidad de este resultado de la muestra

(4) Si σ_x^2 y σ_y^2 son desconocidas y además $\sigma_x^2 \neq \sigma_y^2$.

* $\bar{x} - \bar{y} \rightarrow N\left(\mu_x - \mu_y ; \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}\right)$

Supuestos q. las muestras son suficiente / grandes (>100), pa' q. siga una "Z."

Tipificando: $Z = \dfrac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$

donde $Z \sim N(0;1)$.

DISTRIBUCIONES MUESTRALES PARA r Y $\overline{p}$, n = 3

Variable aleatoria		
Número de éxitos (r)	Proporción de éxitos ($\overline{p}$)	Probabilidad P (r) o P ($\overline{p}$)
0	0.00	0.343
1	0.33	0.294 + 0.147 = 0.441
2	0.67	0.063 + 0.126 = 0.189
3	1.00	0.027
		1.000

(3) Si σ_x^2 y σ_y^2 son conocidas, pero $\sigma_x^2 \neq \sigma_y^2$

$$\overline{x} - \overline{y} \rightarrow N\left(\mu_x - \mu_y ; \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right)$$

Tipificando:

$$z = \frac{\overline{x} - \overline{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

donde $z \sim N(0;1)$

CÁLCULO DE LA MEDIA Y DE LA VARIANZA PARA CADA DISTRIBUCIÓN MUESTRAL

Número de éxitos			Proporción de éxitos		
$r \cdot P(r)$	$(r-\mu_r)^2$	$P(r)\cdot(r-\mu_r)^2$	$\bar{p}\cdot P(\bar{p})$	$(\bar{p}-\mu_{\bar{p}})^2$	$P(\bar{p})\cdot(\bar{p}-\mu_{\bar{p}})^2$
0.000	0.81	0.27783	0.000	0.09000	0.03087
0.441	0.01	0.00441	0.146	0.00111	0.00049
0.378	1.21	0.22869	0.127	0.13444	0.02541
0.081	4.41	0.11907	0.027	0.49000	0.01323
0.900		0.63000	0.300		0.07000
$=\mu_r$		$=\mathrm{Var}(r)$	$=\mu_{\bar{p}}$		$=\mathrm{Var}(\bar{p})$

(2)

CONDICIONES

* Población: $X \to N(\mu_x; ?)$
  $Y \to N(\mu_y; ?)$ independientes.

INDEPENDIENTES

$\sigma_x^2 = \sigma_y^2$ desconocidas

* 2 muestras de tamaño n_x y n_y.
  Podemos calcular S_x^2 y S_y^2.

* $\bar{x}-\bar{y} \to N\left(\mu_x-\mu_y;\ S_P\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}\right)$

donde: $S_P^2 = \frac{[(n_x-1)S_x^2+(n_y-1)S_y^2]}{(n_x+n_y-2)}$

Tipificando

$$T = \frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{S_P\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \to t_{n_x+n_y-2}$$

Distribución Muestral de la Diferencia de dos Medias $\bar{X}-\bar{Y}$

Condiciones

* Población: v.a. $X \to N(\mu_X; \sigma)$
  v.a. $Y \to N(\mu_Y; \sigma)$ Independientes.

  σ^2 conocida

  $\sigma_X^2 = \sigma_Y^2$

* Muestra: elegimos dos muestras

  $n_X \to$ m.a.s de la población X.

  $n_Y \to$ m.a.s de la población Y.

Distribuciones de $(\bar{X}-\bar{Y})$

* Sabemos que $\bar{X} \to N(\mu_X; \sigma^2/n_X)$.

  $\bar{Y} \to N(\mu_Y; \sigma^2/n_Y)$.

$\bar{X}$ e $\bar{Y}$ INDEPENDIENTES

* $\bar{X}$ e $\bar{Y} \to N\left(\mu_X-\mu_Y, \sqrt{\sigma^2/n_X + \sigma^2/n_Y}\right)$

TIPIFICANDO,

$$Z = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}}$$

donde $Z \sim N(0,1)$

Esta total/ prohibido sumar las desviaciones típicas directa/, por ello se utiliza la varianza.

TEMA 3

DISTRIBUCIONES MUESTRALES DE PROBABILIDAD EN EL MUESTREO ALEATORIO SIMPLE

I. Introducción: conceptos previos.

II. Aditividad de variables: parámetros y naturaleza de la distribución resultante.

III. Concepto de distribución muestral.

IV. Distribución muestral de la media.

- A. Parámetros y naturaleza.
- B. Ejemplos.
- C. Caso de varianza poblacional desconocida.
- D. Análisis que involucran a dos poblaciones.

V. Distribución muestral de la varianza.

- A. Conceptos previos.
- B. Parámetros y naturaleza.
- C. Ejemplos.
- D. Análisis que involucran a dos poblaciones.
- E. Ejemplos.

VI. Distribuciones muestrales derivadas de la binomial.

- A. Distribución muestral del número de éxitos.
- B. Distribución muestral de la proporción de éxitos.
- C. Análisis que involucran a dos poblaciones.

I. INTRODUCCION: CONCEPTOS PREVIOS

Una **población** es un conjunto de N valores u observaciones que representan el "estado de naturaleza" de un fenómeno o característica medible X.

→ La edad es una característica que se puede medir (por ejemplo, en años). Las edades en años de todos los habitantes de Guipúzcoa formarían así las N observaciones de una población.

Una **muestra** es un conjunto de n valores u observaciones extraídos de una población de tamaño N (donde $n < N$).

→ Las edades en años de 1.000 habitantes de Guipúzcoa conforman una muestra de tamaño n=1.000 extraída de la población "Edades en años de todos los habitantes de Guipúzcoa".

Los **parámetros** son medidas que resumen la información de todas las observaciones de la población.

→ Algunos parámetros comunes son la media poblacional (μ) y la varianza poblacional (σ^2).

Los parámetros de una población son valores únicos y constantes para esa población en un momento dado.

→ Por ejemplo, la edad media en años de todos los habitantes de Guipúzcoa es el parámetro μ de la población "Edad en años de todos los habitantes de Guipúzcoa". Obviamente la edad media es única, y constante en un momento dado del tiempo.

Los **estadísticos** son medidas que resumen la información de todas las observaciones de la muestra.

→ Algunos estadísticos comunes son la media muestral ($\bar{x}$) y la varianza muestral (S^2).

Los estadísticos, en contraposición a los parámetros, no son valores fijos, ya que varían (o al menos pueden variar) de una muestra a otra.

→ Por ejemplo, si seleccionamos una primera muestra de 1.000 guipuzcoanos y calculamos su edad media, ésta no tiene por qué coincidir con la edad media calculada sobre otra muestra de tamaño 1.000 extraída de la misma población.

Si la muestra se obtiene de modo imparcial y técnicamente adecuado, los estadísticos obtenidos en una muestra pueden proporcionar información útil para realizar una inferencia acerca del verdadero valor de los parámetros.

II. ADITIVIDAD DE VARIABLES: PARAMETROS Y NATURALEZA DE LA DISTRIBUCION RESULTANTE

Sea

$$X_1, X_2, \ldots, X_n$$

un conjunto de n variables aleatorias independientes **normalmente** distribuidas con medias y varianzas:

$$E(X_i) = \mu_i$$

$$Var(X_i) = \sigma_i^2$$

para i=1,2,3,...,n. Si con este conjunto de n variables formamos una combinación lineal, a la que denominaremos Y, del modo

$$Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

en donde

$$a_1, a_2, \ldots, a_n$$

son constantes, se cumplirá que Y es una variable aleatoria con distribución **normal** y media y varianza:

$$E(Y) = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n = \sum_{i=1}^{n} a_i\mu_i$$

$$Var(Y) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_n^2\sigma_n^2 = \sum_{i=1}^{n} a_i^2\sigma_i^2$$

Debe notarse que la hipótesis de normalidad no es necesaria para obtener las fórmulas de la media y de la varianza de Y. De hecho, si

$$X_1, X_2, \ldots, X_n$$

es un conjunto de n variables aleatorias independientes idénticamente distribuidas (VAIID) con medias y varianzas:

$$E(X_i) = \mu_i$$

$$Var(X_i) = \sigma_i^2$$

para i=1,2,3,...,n, entonces para

$$Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

en donde

$$a_1, a_2, \ldots, a_n$$

son constantes se tiene también que:

$$E(Y) = \sum_{i=1}^{n} a_i\mu_i$$

$$Var(Y) = \sum_{i=1}^{n} a_i^2\sigma_i^2$$

En este caso (si la naturaleza de las n variables de partida

$$X_1, X_2, \ldots, X_n$$

no es normal), la combinación lineal Y tendrá naturaleza normal **siempre que el número de variables que la forman tienda a infinito.** Podemos considerar esta afirmación como una aproximación al teorema central del límite.

Por tanto:

Naturaleza de la combinación lineal

$$Si\ X_i \Rightarrow Normal,\ entonces\ Y \Rightarrow Normal,\ \forall\ n$$

$$Si\ X_i \neq Normal,\ entonces\ Y \rightarrow Normal,\ si\ n \rightarrow \infty$$

III. CONCEPTO DE DISTRIBUCION MUESTRAL

Una muestra aleatoria es un conjunto

$$X_1, X_2, \ldots, X_n$$

de n VAIID (variables aleatorias independientes idénticamente distribuídas) con medias y varianzas:

$$E(X_i) = \mu$$

$$Var(X_i) = \sigma^2$$

donde μ y σ^2 son la media y la varianza de la población de la que se extrae la muestra.

Partimos de una población de tamaño N (esto es, con N datos u observaciones de una variable aleatoria X) que sigue una cierta distribución y que está caracterizada por una media poblacional (μ) y una varianza poblacional (σ^2).

De esta población extraemos una primera muestra aleatoria de tamaño n, con lo que tendremos n variables aleatorias independientes idénticamente distribuidas (VAIID):

$$X_1, X_2, \ldots, X_n$$

Con estas n VAIID vamos a calcular un estadístico al que llamaremos Y. Este estadístico será una función de las n observaciones o VAIID:

$$Y = h(X_1, X_2, \ldots, X_n)$$

A continuación, repetimos el proceso y volvemos a extraer una segunda muestra aleatoria de tamaño n, volviendo a calcular el mismo estadístico que antes.

Repetiremos este proceso tomando de la población todas las posibles muestras aleatorias de tamaño n, hasta llegar a la k-ésima muestra, que será la última muestra de tamaño n que se puede extraer de la población.

Evidentemente, podemos decir que el estadístico calculado en cada una de las muestras constituye una nueva variable aleatoria, ya que es función de un conjunto de n VAIID. Al finalizar el proceso, dispondremos de k valores de un estadístico que hemos calculado para cada una de las k muestras, es decir, dispondremos de k valores de la **nueva variable aleatoria** que hemos construido. Estos k valores de la nueva variable aleatoria así construída forman la **distribución muestral**.

> **Podemos por tanto decir que una distribución muestral es una variable aleatoria constituida por el valor de un estadístico calculado en todas las posibles muestras de tamaño n que se pueden extraer de la población de partida.**

Al igual que sucede con toda variable aleatoria podremos preguntarnos por el **valor esperado**, la **varianza** y la **naturaleza** de la distribución muestral.

En función de cuál sea el estadístico calculado en la muestra hablaremos de distribución muestral de la media, distribución muestral de la varianza, etcétera.

IV. DISTRIBUCION MUESTRAL DE LA MEDIA

Supondremos que en la población de partida se estudia una variable aleatoria X que se distribuye normalmente con:

$$E(x) = \mu_x$$
$$Var(x) = \sigma^2{}_x$$

En el caso de la distribución muestral de la media, el estadístico resumen que calcularemos en cada una de las muestras de tamaño n extraídas de la población será la media muestral:

$$\overline{X} = \frac{(X_1 + X_2 + \ldots + X_n)}{n}$$

donde

$X_1 \Rightarrow N(\mu_x; \sigma^2{}_x) \Rightarrow$ *primera observación*

$X_2 \Rightarrow N(\mu_x; \sigma^2{}_x) \Rightarrow$ *segunda observación*

.

.

$X_n \Rightarrow N(\mu_x; \sigma^2{}_x) \Rightarrow$ *enésima observación*

IV.A. Parámetros y naturaleza

Se puede observar que el estadístico media muestral es una combinación lineal de las n VAIID que forman la muestra, y que como tal combinación lineal constituye una nueva variable aleatoria. Esta nueva variable aleatoria seguirá necesariamente algún tipo de distribución, tendrá un valor esperado y tendrá una varianza.

El valor esperado de la combinación lineal media muestral es:

$$E[\overline{X}] = E[\frac{X_1 + X_2 + \ldots + X_n}{n}] =$$

$$= \frac{1}{n}E[X_1] + \frac{1}{n}E[X_2] + \ldots + \frac{1}{n}E[X_n] =$$

$$= \frac{1}{n}\mu_x + \frac{1}{n}\mu_x + \ldots + \frac{1}{n}\mu_x = \frac{n\,\mu_x}{n} = \mu_x$$

$$E[\overline{X}] = \mu_{\overline{x}} = \mu_x$$

El valor esperado de la distribución muestral de medias es igual a la media poblacional.

Esta propiedad va a cumplirse siempre, independientemente del tamaño de la muestra con el que se construya la distribución muestral y de la naturaleza con que se distribuya la variable X en la población de partida.

La varianza de la combinación lineal media muestral es:

$$Var(\overline{X}) = Var\frac{X_1+X_2+\ldots+X_n}{n} =$$

$$= \frac{1}{n^2}Var(X_1)+\frac{1}{n^2}Var(X_2)+\ldots+\frac{1}{n^2}Var(X_n) =$$

$$= \frac{1}{n^2}\sigma^2{}_x+\frac{1}{n^2}\sigma^2{}_x+\ldots+\frac{1}{n^2}\sigma^2{}_x = \frac{n\,\sigma^2{}_x}{n^2} = \frac{\sigma^2{}_x}{n}$$

Si el muestreo es sin reemplazamiento:

$$Var(\overline{X}) = \frac{\sigma^2{}_x}{n}\left(\frac{N-n}{N-1}\right)$$

$\left(\frac{N-n}{N-1}\right) \Rightarrow$ *Corrector de poblaciones finitas*

si $N \to \infty$ *entonces CPF* $\to 1$

Es decir, la varianza de la distribución muestral de medias es la varianza de la población de partida dividida entre el tamaño de muestra sobre el que se calcula la distribución muestral. En casos de muestreo sin reemplazamiento (MSR) hay que introducir un corrector de poblaciones finitas, cuyo valor es despreciable si el tamaño de la población de partida es lo suficientemente grande.

A la desviación estándar de la distribución muestral se la denomina error estándar

Sobre la **naturaleza de la combinación lineal media muestral** podemos señalar que, en el caso de que las n VAIID que componen la muestra sean de naturaleza normal, esto es, en el caso de que la población de partida sea de naturaleza normal, la combinación lineal media muestral tendrá naturaleza normal, independientemente del tamaño de la muestra elegida.

En el caso de que las n VAIID que componen la muestra no sean de naturaleza normal, o lo que es lo mismo, en el caso de que la población de la que se ha extraído la muestra no sea de naturaleza normal, será de aplicación el teorema central del límite que dice:

> **"Independientemente de la naturaleza de la distribución de la población, la distribución muestral de medias tendrá naturaleza normal siempre que el tamaño de la muestra tienda a infinito".**

Una cuestión obvia que se plantea es cómo de grande debe ser el tamaño de la muestra para que pueda ser considerado "infinito". En la práctica se considera "suficientemente grande" como para garantizar el cumplimiento del TCL todo tamaño muestral de al menos 30 observaciones.

IV.B. Ejemplos

Para clarificar el concepto de distribución muestral de la media, expondremos tres ejemplos:

Ejemplo 1: Una población consta de los números 2,3,6,8 y 11. Consideremos todas las posibles muestras de tamaño 2 que pueden tomarse **con reemplazamiento** de esa población. Hallar la media de la población, la desviación estándar de la población, la media de la distribución muestral de medias y el error estándar de la distribución muestral de medias. Calcular de nuevo todos los valores para el caso de muestreo sin reemplazamiento.

a) La media de la población será:

$$\mu = \frac{2+3+6+8+11}{5} = 6$$

b) La desviación estándar de la población será:

$$\sigma = \sqrt{\sigma^2} =$$

$$= \sqrt{\frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5}} =$$

$$= \sqrt{\frac{54}{5}} = 3.29$$

c) Existen 25 posibles muestras de tamaño 2 que se pueden extraer de una población de tamaño N=5. Las posibles muestras y su media son:

(2,2)	--> 2	(2,3)	--> 2.5
(2,6)	--> 4	(2,8)	--> 5
(2,11)	--> 6.5	(3,2)	--> 2.5
(3,3)	--> 3	(3,6)	--> 4.5
(3,8)	--> 5.5	(3,11)	--> 7
(6,2)	--> 4	(6,3)	--> 4.5
(6,6)	--> 6	(6,8)	--> 7
(6,11)	--> 8.5	(8,2)	--> 5
(8,3)	--> 5.5	(8,6)	--> 7
(8,8)	--> 8	(8,11)	--> 9.5
(11,2)	--> 6.5	(11,3)	--> 7
(11,6)	--> 8.5	(11,8)	--> 9.5
(11,11)	--> 11		

La media de la distribución muestral de medias será:

$$\mu_{\bar{X}} = \frac{2+2.5+4+5+6.5+...+11}{25} = \frac{150}{25} = 6$$

$$\mu_{\bar{X}} = \mu_X = 6$$

El error estándar de la distribución muestral de medias será:

$$\sigma_{\bar{x}} = \sqrt{\frac{(2-6)^2+(2.5-6)^2+\ldots+(11-6)^2}{25}} =$$

$$= 2.32$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{3.29}{\sqrt{2}} = 2.32$$

d) En el caso de muestreo sin reemplazamiento tanto la media poblacional como la varianza poblacional seguirán valiendo 6 y 3.29 respectivamente, ya que estos parámetros no se ven afectados por el método de muestreo.

El conjunto de muestras de tamaño 2 que se pueden extraer sin reemplazamiento de esta población de tamaño 5 son 10, a saber:

(2,3)	--> 2.5	(2,6)	--> 4
(2,8)	--> 5	(2,11)	--> 6.5
(3,6)	--> 4.5	(3,8)	--> 5.5
(3,11)	--> 7	(6,8)	--> 7
(6,11)	--> 8.5	(8,11)	--> 9.5

La media de la distribución muestral de medias será:

$$\mu_{\bar{x}} = \frac{2.5+4+5+6.5+4.5+\ldots+9.5}{10} = 6$$

$$\mu_{\bar{x}} = \mu_x = 6$$

El error estándar de la distribución muestral de medias será:

$$\sigma_{\bar{x}} = \sqrt{\frac{(2.5-6)^2+(4-6)^2+\ldots+(9.5-6)^2}{10}} = 2.0$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{3.29}{\sqrt{2}}\sqrt{\frac{5-2}{5-1}} = 2.0$$

Ejemplo 2: Se tiene una máquina para llenar por término medio 500 gramos de cereal en cajas de cartón. Supóngase que la cantidad de cereal que se deposita en cada caja es una variable aleatoria normal de media 500 y de desviación estándar 20. Como parte del proceso de calidad se toma periódicamente una muestra de 25 cajas llenas y se pesa su contenido. Si el peso medio de la muestra es superior a 510 gramos o inferior a 490 gramos se detiene el proceso de llenado de la máquina. Calcular la probabilidad de que en uno de estos procesos de control de calidad se detenga el proceso de llenado.

X: Cantidad de cereal medido en gramos que la máquina deposita en 1 caja.

μ_x= 500 gramos σ_x = 20 gramos X -> NORMAL

Extraemos una muestra de 25 cajas, con lo que n=25. Cada una de las observaciones constituirá una variable aleatoria que en principio, por tratarse de un muestreo sin reemplazamiento no podremos calificar de independientes. Sin embargo, podemos considerar que la población de la cual se extrae la muestra de 25 cajas es lo suficientemente grande como para considerarla infinita. Ello nos permite considerar a las 25 variables constituidas por las 25 observaciones como variables independientes a todos los efectos.

X_1: Peso del cereal contenido en la 1ª caja
X_2: Peso del cereal contenido en la 2ª caja
.....
X_n: Peso del cereal contenido en la enésima caja

Cada una de estas 25 variables aleatorias seguirá una distribución normal con media 500 y desviación estándar 20. Por ello podemos decir que son 25 variables aleatorias independientes e idénticamente distribuidas.

A continuación definimos una nueva variable a la que llamaremos $\overline{X}_{25}$ y que definiremos como "peso medio del cereal contenido en una muestra de 25 cajas":

$$\overline{X}_{25} = \frac{X_1 + X_2 + \ldots + X_{25}}{25}$$

$$E[\overline{X}_{25}] = \mu_X = 500 \ gramos$$

$$\sigma^2_{\overline{X}_{25}} = \frac{\sigma^2_x}{25}\frac{N-n}{N-1} =$$

$$= \frac{400}{25} = 16\ gramos^2$$

Conocemos tanto la media como la varianza de la distribución muestral de la media. Sólo nos resta por conocer su naturaleza.

Tal y como se dice en el enunciado la población de partida tiene naturaleza normal. En ese caso, la combinación lineal media muestral tiene también naturaleza normal independientemente del tamaño de la muestra.

Por todo ello sabemos:

$\overline{X}_{25}$: Peso medio del cereal en una muestra de 25 cajas.

$E[\overline{X}_{25}]=500$ gramos Var $[\overline{X}_{25}]=16$ gr ^{2}

Naturaleza $\overline{X}_{25}$ --> NORMAL

La probabilidad de que el proceso se detenga puede expresarse del siguiente modo:

$$P[\overline{X}_{25}>510]+P[\overline{X}_{25}<490] =$$

$$=1-P[490\leq\overline{X}_{25}\leq510] =$$

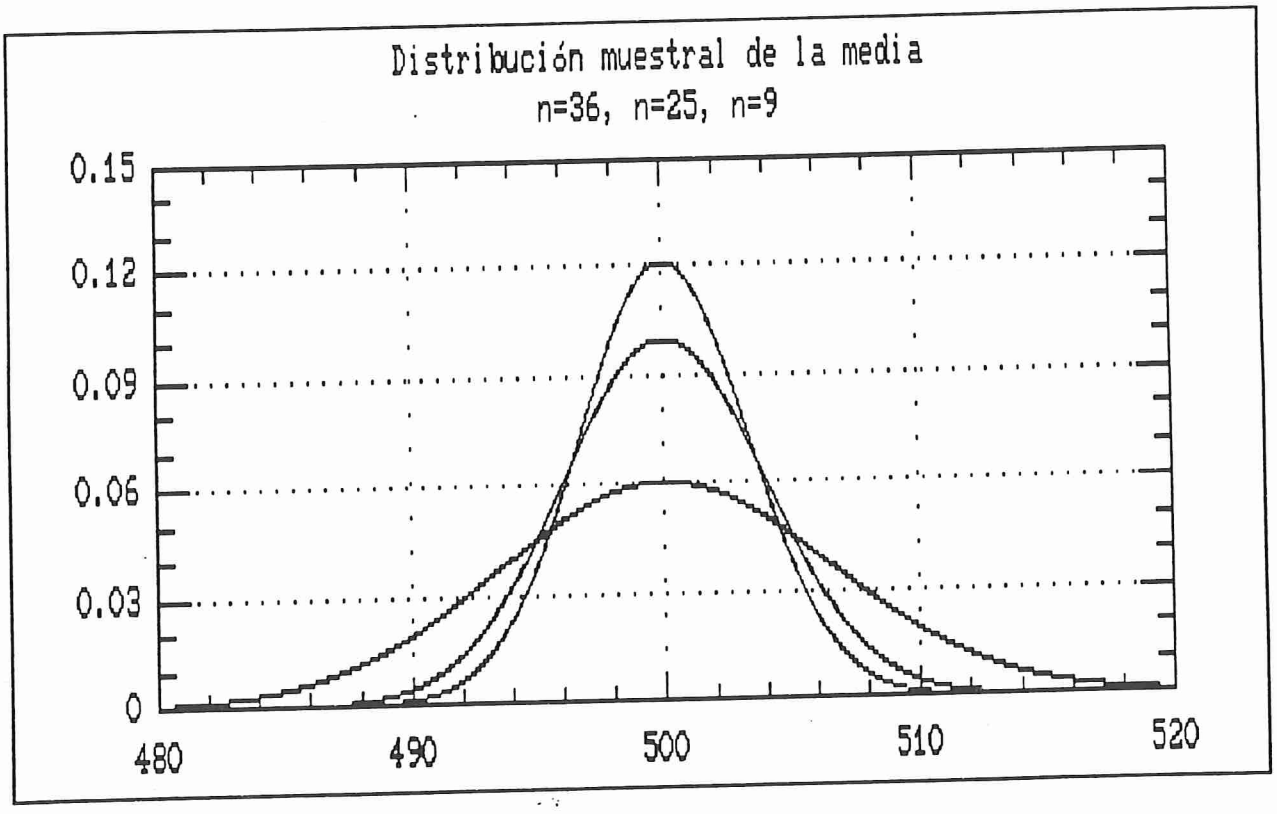
$$=1-P[\frac{490-500}{4}\leq Z\leq\frac{510-500}{4}] =$$

$$=1-P[-2.5\leq Z\leq2.5]=0.0124$$

Esto significa que aproximadamente en 1 de cada 100 ocasiones en que se realice el control de calidad tomando muestras de 25 cajas y analizando su peso medio en cereal, se detendrá el proceso.

Podemos comprobar que en el caso de haber tomado muestras de 9 cajas, la probabilidad sería de 0,1336, mientras que en el caso de tomar muestras de 36 cajas, sería de 0,0027.

Esto nos lleva a pensar que cuanto más dispersa es una distribución muestral, esto es, cuanto mayor sea su varianza, tanto más probable será la aparición de valores extremos. Por el contrario, cuanto más concentrada sea una distribución muestral, o lo que es lo mismo, cuanto menor sea su varianza, tanto menos probable será la aparición de valores extremos.



Como se desprende de la expresión de la varianza de la distribución muestral de medias, a medida que el tamaño de la muestra elegida aumenta, dicha varianza disminuye. Por el contrario, cuanto menor sea el tamaño de la muestra elegida, la varianza de la distribución muestral de medias será mayor.

Esto puede percibirse de modo intuitivo. En el caso de que el tamaño de la muestra sea el mínimo posible, esto es, una muestra de tamaño 1, el número k de muestras que se podrá extraer de la población será igual al tamaño de la población. En ese caso, la distribución muestral de la media será idéntica a la distribución de la población, por lo que la varianza de la distribución muestral de la media será idéntica a la varianza de la distribución de la población.

Por el contrario, en el caso de que el tamaño de la muestra sea el máximo posible, esto es, una muestra de tamaño n igual al tamaño N de la población, el número k de muestras que se podrá extraer de la población será igual a 1 y por tanto la varianza de la distribución muestral de medias será nula.

$$\sigma^2_{\bar{x}} = \frac{\sigma^2_x}{n}\left(\frac{N-n}{N-1}\right)$$

$$si\ n = 1 \Rightarrow \sigma^2_{\bar{x}} = \frac{\sigma^2_x}{1}\left(\frac{N-1}{N-1}\right) = \sigma^2_x$$

$$si\ n = N \Rightarrow \sigma^2_{\bar{x}} = \frac{\sigma^2_x}{N}\left(\frac{N-N}{N-1}\right) = 0$$

Ejemplo 3: Supóngase que el número de barriles de petróleo que produce diariamente un pozo es una variable aleatoria de naturaleza no especificada. Si se observa la producción de 64 días seleccionados en forma aleatoria y si se sabe que la desviación estándar del número de barriles producidos por el pozo es de 16, calcular la probabilidad de que la media muestral se encuentre a no más de cuatro barriles de la producción media por día.

X: Número de barriles de petróleo que produce el pozo en 1 día.

μ_x= ? σ_x= 16 barriles Naturaleza X --> ?

Consideramos una muestra de 64 días, con lo que tendremos 64 variables aleatorias.

X_1 : Número de barriles producidos el primer día
X_2 : Número de barriles producidos el segundo día
.....
X_{64} : Número de barriles producidos el día n

Estas 64 variables estarán idénticamente distribuidas. Esto es, su media será igual a la media de la población y su varianza será igual a la varianza de la población. Del mismo modo, su naturaleza será idéntica a la de la población, aunque ésta, al igual que la media poblacional, sea desconocida. Por otra parte, a pesar de que el muestreo se lleve a cabo sin reemplazamiento (por razones obvias no podemos reemplazar un día), el tamaño de la población (que puede ser considerado infinito), nos permite considerar a las 64 variables definidas como variables independientes.

A continuación definimos una nueva variable aleatoria $\overline{X}_{64}$, que definiremos como "número medio de barriles que produce el pozo en una muestra de 64 días".

$$\overline{X}_{64} = \frac{X_1 + X_2 + \ldots + X_{64}}{64}$$

$$E[\overline{X}_{64}] = \mu_x = ?$$

$$Var[\overline{X}_{64}] = \frac{\sigma^2_x}{n}\frac{N-n}{N-1} = \frac{16^2}{64} = 4$$

$$Nat(\overline{X}_{64}) \Rightarrow NORMAL \Rightarrow TCL \ (n = 64 > 30)$$

La importancia de este ejemplo recae sobre el hecho de que a pesar de que la población de partida presenta una distribución desconocida, podemos considerar que la distribución muestral de la media sigue una distribución normal gracias al teorema central del límite.

Por lo demás, la probabilidad pedida puede expresarse del siguiente modo:

$$P[\,|\overline{X}_{64} - \mu_X| \leq 4\,] = P[-4 \leq \overline{X}_{64} - \mu_X \leq 4\,] =$$

$$= P[-2 \leq Z \leq 2\,] = 0.9544$$

Este resultado se puede interpretar del siguiente modo. En 95 de cada 100 veces que tomemos muestras de 64 días de la producción del pozo, la media muestral se separará en menos de cuatro barriles de la media real de la población, aunque ésta sea desconocida.

V. DISTRIBUCION MUESTRAL DE LA VARIANZA

El estadístico resumen que calcularemos en las muestras de tamaño n que extraemos de la población será la varianza muestral (S^2), que mide la dispersión entre las observaciones. El significado de S^2 para formular inferencias de σ^2 es comparable con el que tiene $\bar{x}$ para formular inferencias con respecto a μ.

V.A. Conceptos previos

a) Sea un conjunto de n variables Z, esto es, de naturaleza normal con valor esperado 0 y con varianza 1.

La suma de los cuadrados de esas n variables Z dará como resultado una nueva variable aleatoria: una Chi-cuadrado con n grados de libertad.

$$Si \quad Z_i \sim N(0;1) \quad entonces:$$

$$Z_1^2 + Z_2^2 + \ldots + Z_n^2 = \chi_n^2$$

$$\sum_{i=1}^{n} Z_i^2 = \chi_n^2$$

$$Si \quad X_i \sim N(\mu_x;\sigma_x^2) \quad entonces:$$

$$\left(\frac{X_1-\mu_x}{\sigma_x}\right)^2 + \left(\frac{X_2-\mu_x}{\sigma_x}\right)^2 + \ldots + \left(\frac{X_n-\mu_x}{\sigma_x}\right)^2 = \chi_n^2$$

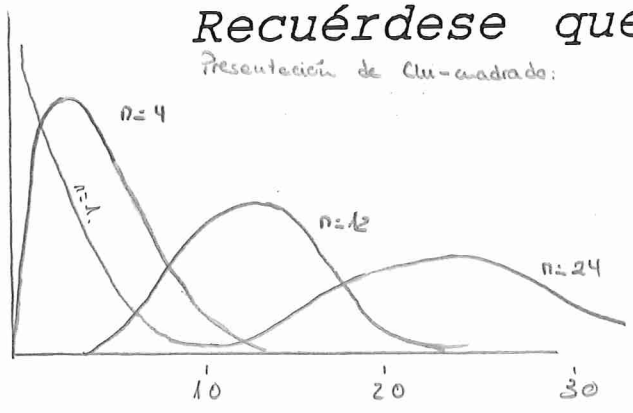
$$\sum_{i=1}^{n} \left(\frac{X_i-\mu_x}{\sigma_x}\right)^2 = \chi_n^2$$

$$Recuérdese \ que: \quad E(\chi_n^2) = n$$

$$Var(\chi_n^2) = 2n$$

El nº de grados de libertad.

Presentación de Chi-cuadrado:



b) Si X_1 y X_2 son dos variables aleatorias independientes y cada una tiene una distribución chi-cuadrada con v_1 y v_2 grados de libertad respectivamente, entonces una nueva variable aleatoria del tipo $Y = X_1 + X_2$ tendrá una distribución chi-cuadrada con $v_1 + v_2$ grados de libertad.

V.B. Parámetros y naturaleza

Partimos de una población en la que estudiamos una variable X que necesariamente debe ser de **naturaleza normal**, con media μ_x y desviación estándar σ_x.

De esta población extraemos una muestra aleatoria de tamaño n, que estará formada por n VAIID, todas ellas de naturaleza normal y con media y desviación estándar idénticas a la media y desviación estándar poblacionales.

$$X_1 , X_2 , \ldots , X_n$$

$$X_i \sim N(\mu_x ; \sigma_x^2)$$

En esta muestra que hemos extraído de la población calculamos la **varianza muestral (S^2)** cuya expresión es la siguiente:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

→ tamaño de la muestra.

→ cuasivarianza muestral.

* la varianza no puede ser negativa. Siempre será (+), y por lo tanto no puede seguir la distribución normal.

Operando con esta expresión de la varianza muestral llegaremos a otra expresión, basada en la distribución chi-cuadrado, que nos servirá para resolver preguntas probabilísticas sobre la distribución muestral de varianzas y para hacer inferencias sobre el valor del parámetro varianza poblacional.

Tenemos así que:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} \rightarrow (n-1)\,S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} (X_i - \mu_x - \overline{X} + \mu_x)^2 =$$

$$= \sum_{i=1}^{n} [\,(X_i - \mu_x) - (\overline{X} - \mu_x)\,]^2 =$$

$$= \sum_{i=1}^{n} [\,(X_i - \mu_x)^2 - 2\,(X_i - \mu_x)\,(\overline{X} - \mu_x) + (\overline{X} - \mu_x)^2\,] =$$

$$= \sum_{i=1}^{n} (X_i - \mu_x)^2 - 2\,(\overline{X} - \mu_x) \sum_{i=1}^{n} (X_i - \mu_x) + n\,(\overline{X} - \mu_x)^2 =$$

$$= \sum_{i=1}^{n} (X_i - \mu_x)^2 - 2\,(\overline{X} - \mu_x)\,n\,(\overline{X} - \mu_x) + n\,(\overline{X} - \mu_x)^2 =$$

$$= \sum_{i=1}^{n} (X_i - \mu_x)^2 - n\,(\overline{X} - \mu_x)^2$$

$$(n-1)\,S^2 + n\,(\overline{X} - \mu_x)^2 = \sum_{i=1}^{n} (X_i - \mu_x)^2$$

Dividiendo las tres expresiones entre σ^2

$$\frac{(n-1)\,S^2}{\sigma^2} + \frac{n\,(\overline{X} - \mu_x)^2}{\sigma^2} = \frac{\sum_{i=1}^{n} (X_i - \mu_x)^2}{\sigma^2}$$

$$\chi^2_{n-1} + \chi^2_1 = \chi^2_n$$

De donde deducimos que $\dfrac{(n-1)\,S^2}{\sigma^2} = \chi^2_{n-1}$

¿Cuáles son entonces el valor esperado y la varianza de la distribución muestral de varianzas?

$$\frac{(n-1)S^2}{\sigma^2} = \chi^2_{n-1}$$

$$E(S^2) = E\left(\frac{\chi^2_{n-1}\sigma^2}{n-1}\right) = \frac{\sigma^2}{n-1}E(\chi^2_{n-1})$$

$$= \frac{\sigma^2}{n-1}(n-1) = \sigma^2$$

$$Var(S^2) = Var\left(\frac{\chi^2_{n-1}\sigma^2}{n-1}\right) = \frac{\sigma^4}{(n-1)^2}Var(\chi^2_{n-1})$$

$$= 2\frac{\sigma^4}{(n-1)^2}(n-1) = \frac{2\sigma^4}{n-1}$$

V.C. Ejemplos

Ejemplo 1: La longitud de las piezas fabricadas en un proceso productivo es una variable aleatoria normal con media igual a 10 cm. y desviación estándar 0,1 cm. Se toma para medir una muestra de tamaño 25 del proceso de fabricación. ¿Cuál es la probabilidad de que el valor de la varianza muestral sea superior a 0,014?

X: Longitud de 1 pieza del proceso productivo

$\mu_x = 10$; $\sigma_x = 0,1$; Naturaleza X --> NORMAL

Mediante muestreo sin reemplazamiento extraemos una muestra de tamaño 25. Dado que la población es infinita, podemos considerar las veinticinco mediciones de la muestra como 25 VAIID.

$$X_1 , X_2 , \ldots , X_{25}$$

$$X_i \sim N(10; 0.1^2)$$

A continuación, definimos una nueva variable que designaremos por S^2_{25} y que definiremos como "Varianza de la longitud de las piezas fabricadas en una muestra de 25 piezas".

$$S^2_{25} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{24}$$

La probabilidad pedida se puede calcular del siguiente modo:

$$P[S^2_{25} > 0.014] = P\left(\frac{(n-1)S^2_{25}}{\sigma^2} > \frac{(25-1)0.014}{0.1^2}\right)$$

$$= P[\chi^2_{24} > 33.6] \approx 0.1$$

Este resultado se puede interpretar del siguiente modo: aproximadamente 1 de cada 10 veces en que analicemos muestras de 25 piezas la dispersión total de las mismas, medida en términos de varianza, será superior a 0,014.

Ejemplo 2: El Ayuntamiento de una ciudad ha recibido un informe de la agrupación de constructores de viviendas en el que se afirma que la dispersión en el precio de los alquileres de dichas viviendas, medido en términos de varianza, es de 1 um². El responsable de estudios del Ayuntamiento selecciona al azar siete de dichas viviendas y obtiene los siguientes precios de los alquileres (en um): 100; 102; 104; 100; 106; 104; 98. ¿Qué podemos concluir sobre la afirmación de los constructores de viviendas en ese informe?

X: Precio del alquiler de 1 vivienda construída

Naturaleza de X: tendremos que suponer que X tiene naturaleza **NORMAL**. Aunque el enunciado no diga que la naturaleza de la variable que se estudia en la población es normal, el modelo que se ha desarrollado exige la normalidad de dicha variable.

μ_x: La media poblacional es desconocida.

σ_x : Por el momento, aceptamos como bueno el valor proporcionado por la asociación de constructores. La desviación estándar del precio del alquiler es 1 um.

A través de muestreo sin reemplazamiento tomamos una muestra de 7 viviendas. Dado que el tamaño de la población se considera infinito podemos considerar las 7 observaciones cómo 7 VAIID cuya naturaleza será normal, cuyo valor esperado será desconocido e igual a μ_x y cuya varianza será igual a 1 um².

A continuación calculamos una nueva variable que designaremos S^2_7 y que definiremos como "Varianza del importe del alquiler en una muestra de siete viviendas".

$$S^2_7 = \frac{\sum_{i=1}^{7} (X_i - \overline{X})^2}{6}$$

$$\frac{(100-102)^2}{6} + \frac{(102-102)^2}{6} + \frac{(104-102)^2}{6} + \frac{(100-102)^2}{6} + \frac{(106-102)^2}{6} + \frac{(104-102)^2}{6} + \frac{(98-102)^2}{6}$$

En concreto, en la muestra que hemos extraído de siete viviendas, S^2_7 vale 2.828². La pregunta que nos planteamos entonces es la siguiente **¿Es compatible este valor obtenido en la muestra con la afirmación de la agrupación de constructores de que el verdadero valor de la varianza poblacional en el precio del alquiler es de 1 um²?** Para ello procederemos como sigue:

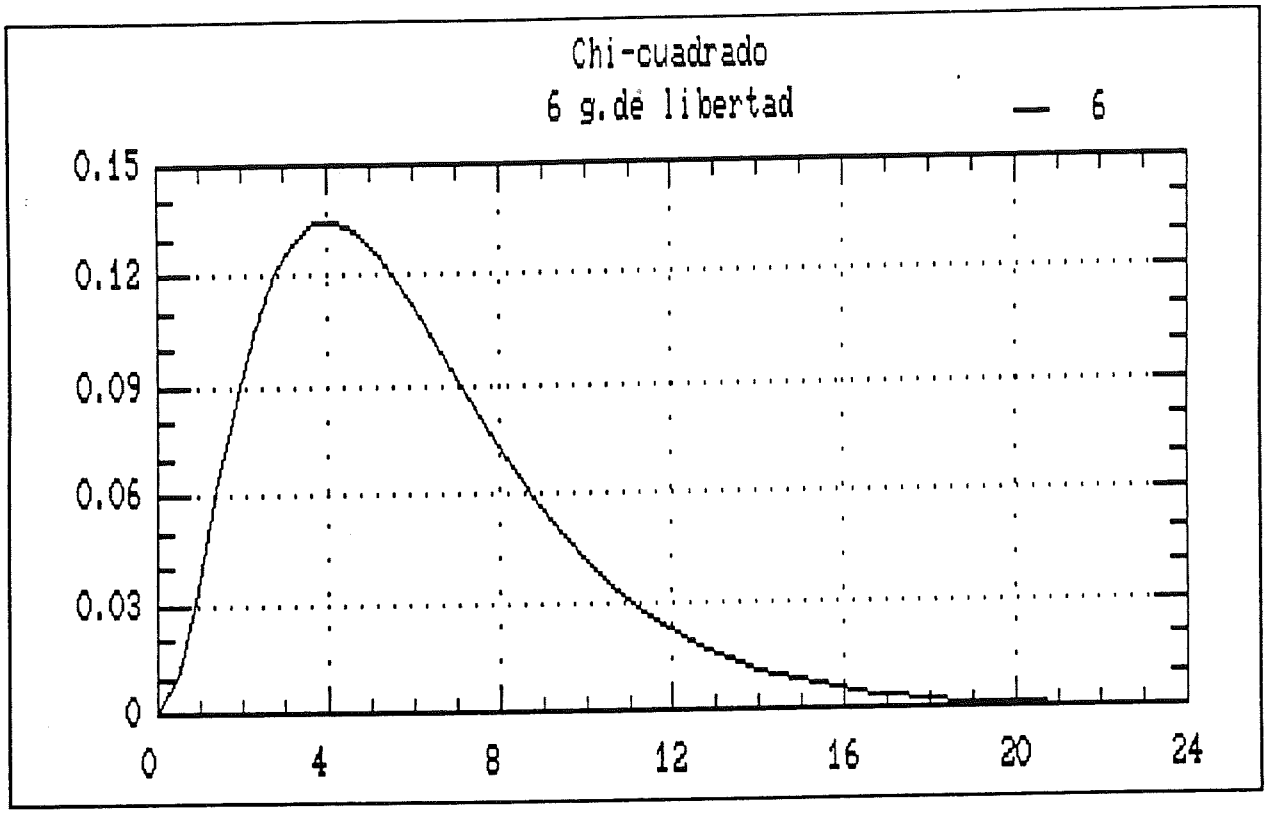
1º) En base al valor de S^2_7 hallado en la muestra, obtener el valor de la chi-cuadrado de 6 grados de libertad:

$$Valor\ de\ \chi^2_6 = \frac{(7-1)2.828^2}{1^2} = 48.05$$

2º) Graficar la chi-cuadrado de 6 grados de libertad:

3º) **Decidir subjetivamente un área central** (que suele ser del 90%, del 95% o del 99%) En este caso, elegiremos un 95% Los valores que encierran un área central del 95% son 1,237 y 14,449.

4º) Situar en la gráfica la puntuación chi-cuadrado de 6 grados de libertad obtenida a partir de la varianza hallada en la muestra extraída. En nuestro caso, el valor 48,05 queda claramente fuera del área central que hemos definido subjetivamente.



5º) Interpretar el resultado: Si la puntuación chi-cuadrada calculada a partir de la varianza hallada en la muestra hubiera quedado incluída dentro del área central subjetivamente decidida diríamos que **la información de la muestra corroboraría los supuestos de partida.**

Sin embargo, en nuestro caso, el valor ha quedado fuera del área central. **Esto significa que la información de la muestra no es compatible con el supuesto de partida, (varianza poblacional igual a 1).** Tenemos por tanto pruebas de que la afirmación de los constructores parece no ser cierta.

Además, como la puntuación chi-cuadrada queda fuera por la derecha, podemos afirmar que el valor de partida (varianza poblacional igual a 1) está infravalorado. Dado que hemos elegido un área del 95% la probabilidad de que la decisión tomada (afirmar que el supuesto de partida es falso) sea errónea será del 5% Podemos decir que el 95% de las veces que analicemos muestras de 7 viviendas y en base a la varianza muestral rechace el supuesto de partida sobre el valor de la varianza poblacional, las decisiones serán acertadas, mientras que en el otro 5% de las ocasiones las decisiones serán erróneas. A este 5% se le llama **nivel de significación** y el 95% (área central subjetivamente elegida) recibe el nombre de **nivel de confianza.**

IV.C. Distribución muestral de la media.
Caso de varianza poblacional desconocida.

Nuestro punto de partida es una variable aleatoria X

$$X \sim NORMAL(\mu_x; \sigma^2_x)$$

En este caso, y como ya sabemos, la distribución muestral de medias estará caracterizada del siguiente modo:

$$\overline{X} \sim NORMAL(\mu_{\overline{x}} = \mu_x \; ; \; \sigma^2_{\overline{x}} = \frac{\sigma^2_x}{n})$$

$$Por\ tanto\ \frac{\overline{X} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}} \sim Z$$

Sin embargo, es muy posible que cuando estemos tratando de realizar inferencias sobre el verdadero valor de la media poblacional **desconozcamos también el verdadero valor de la varianza poblacional.** La solución lógica que adoptaremos será sustituir (estimar) el parámetro desconocido varianza poblacional por un estadístico que podremos calcular en la muestra: la varianza muestral. El hecho de sustituir la varianza poblacional por su estimador (la varianza muestral) hace que la tipificación de la distribución muestral de medias no dé como resultado una Z sino una **t Student con n-1 grados de libertad.** da "t Student se diferencia de la curva de Gauss en el apuntamiento"

$$Si\ \sigma^2_x = ?\ entonces\ \hat{\sigma}^2_x = s^2$$

$$\frac{\overline{x} - \mu_x}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

Algunas propiedades de la t Student son las siguientes:

$$t_\nu = \frac{Z}{\sqrt{(\chi^2_\nu / \nu)}}$$

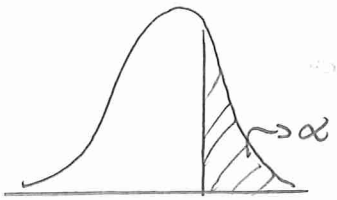
$$t_\nu \ es\ simétrica$$

$$E[t_\nu] = 0 \qquad Var(t_\nu) = \frac{\nu}{(\nu - 2)}$$

$Z \sim N(0,1)$

$X \sim \chi^2_n$

$\frac{Z \cdot \sqrt{n}}{\chi^2_n} = t$



$P[t \geq t_{\alpha,n}] = \alpha$; $P[t \leq t_{\alpha,n}]$ $t_{\alpha,n}$ → Pto. crítico para 1 nivel de significación α.

Demostremos ahora que la tipificación de la distribución muestral de medias cuando se sustituye la varianza poblacional desconocida por su estimación de la varianza muestral no da como resultado una variable Z sino una t Student con n-1 grados de libertad.

$$X \approx N(\mu_x; \sigma_x^2)$$

$$\overline{X} \approx N(\mu_x; \frac{\sigma_x^2}{n}) \Rightarrow \frac{\overline{X}-\mu_x}{\frac{\sigma_x}{\sqrt{n}}} = Z$$

$$Sabemos\ que: \frac{(n-1)S^2}{\sigma_x^2} = \chi^2_{n-1}$$

$$t_{n-1} = \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} = \frac{\frac{\overline{X}-\mu_x}{\frac{\sigma_x}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma_x^2}}{n-1}}} =$$

$$\frac{\frac{\overline{X}-\mu_x}{\frac{\sigma_x}{\sqrt{n}}}}{\frac{\frac{S}{\sigma_x}}{}} = \frac{\overline{X}-\mu_x}{\frac{S}{\sqrt{n}}}$$

La expresión obtenida para la t Student es válida sólo para los casos de muestreo con reemplazamiento. En los casos de muestreo sin reemplazamiento la expresión debe ser corregida con un factor corrector:

$$t_{n-1} = \frac{\overline{X} - \mu_x}{\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}}$$

La t Student la utilizaremos para llevar a cabo **inferencias sobre la media de la población cuando la varianza poblacional sea desconocida.**

Ejemplo: El Departamento de Protección del Medio Ambiente (DPMA) asegura que, en un modelo de automóvil, el consumo de gasolina en carretera es de 1 galón por cada 45 millas. Una organización independiente de consumidores adquiere uno de estos automóviles y lo somete a prueba con el propósito de verificar la cifra proporcionada por el DPMA. El automóvil recorrió una distancia de 100 millas en 25 ocasiones. En cada recorrido se anotó el número de galones necesarios para realizar el viaje. Los 25 ensayos arrojaron un valor medio de 43,5 millas/galón y una desviación estándar de 2,5 millas/galón. Si se supone que el número de millas que se recorre por galón es una variable aleatoria distribuida normalmente, ¿existe con base en esta prueba alguna razón para dudar del dato proporcionado por el DPMA?

X: Número de millas/galón recorridas por el automóvil en 1 viaje.

μ_x = 45 m/g Var(X) = ? Nat (X) ---> NORMAL

A partir de la población formada por el conjunto de viajes que pueden realizar los automóviles de este tipo (a la que podemos considerar infinita), seleccionamos mediante MSR una muestra de 25 viajes. A pesar de que el muestreo se lleve a cabo sin reemplazamiento, el hecho de que podamos considerar la población como infinita hace que las 25 variables aleatorias formadas por el conjunto de los 25 ensayos puedan ser consideradas como 25 VAIID, esto es, normales con media y varianza iguales a la variable de partida X.

A continuación definimos una nueva variable aleatoria que será el "número medio de millas/galón recorridos en una muestra de 25 ensayos" y que designaremos por $\overline{X}_{25}$.

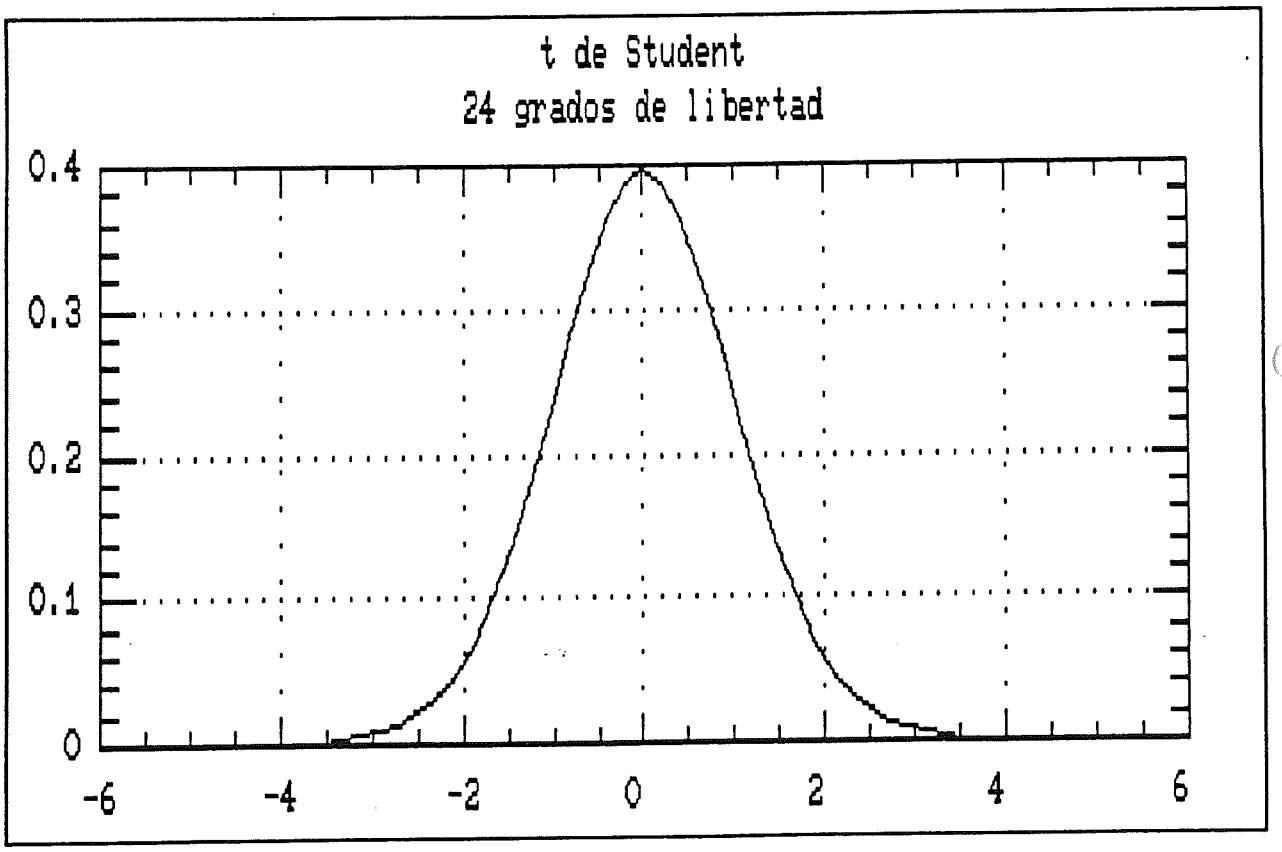
En concreto, en la muestra que hemos elegido, el valor que toma $\overline{X}_{25}$ es 43,5 y el valor de la varianza muestral S^2_{25} es $2,5^2$.

La cuestión a resolver es: ¿Resulta compatible el valor observado en la muestra ($\bar{X}_{25} = 43,5$ m/g) con el valor que asegura el DPMA ($\mu_{\bar{x}} = 45$ m/g)? **¿Son diferencias que se deben al azar exclusivamente o es que realmente resultan ser significativas como para afirmar que el dato suministrado por el DPMA parece no ser cierto?.** Para solucionar esta cuestión los pasos a seguir son:

1º) Obtener la puntuación t_{n-1} a partir de la media y de la desviación estándar muestrales:

$$t_{24} = \frac{43.5-45}{\frac{2.5}{\sqrt{25}}} = -3$$

2º) Representar la t Student de 24 grados de libertad y seleccionar un área central arbitraria (en nuestro caso seleccionaremos un 95%):



Los valores que delimitan un área central del 95% son 2,063 y -2,063. Dado que la puntuación t Student que obtenemos a partir de la muestra es -3 y que este valor cae fuera del área central arbitrariamente seleccionada, debemos concluir que el valor que asegura el DPMA es falso, esto es, que la media poblacional parece no valer 45 m/g. Además, por caer a la izquierda del área central, podemos añadir que el valor que propone el DPMA parece estar sobrevalorado. Esta afirmación que hacemos es cierta con una probabilidad del 95% Esto es, al afirmar que la media poblacional no es 45 millas/galón podemos equivocarnos. Puede suceder (de hecho pasará el 5% de las veces que analicemos muestras de 25 viajes) que aseguremos que la media poblacional no es 45 millas/galón y nos estemos equivocando. Sin embargo, no podemos saber si en este caso concreto nuestra afirmación pertenece al 95% de "afirmaciones correctas" o al 5% de "afirmaciones falsas".

V.D. Distribución muestral de la varianza: análisis que involucran a dos poblaciones.

Recordemos previamente algunas propiedades de la distribución F Fisher:

$$F_{\nu_1,\nu_2} = \frac{\frac{\chi^2_{\nu_1}}{\nu_1}}{\frac{\chi^2_{\nu_2}}{\nu_2}}$$

$$E(F_{\nu_1,\nu_2}) = \frac{\nu_2}{\nu_2-2} \quad para\ \nu_2>2$$

$$Var(F_{\nu_1,\nu_2}) = \frac{\nu_2^2(2\nu_2+2\nu_1-4)}{\nu_1(\nu_2-2)^2(\nu_2-4)} \quad para\ \nu_2>4$$

$$Dominio\ de\ F_{\nu_1,\nu_2}\ es\ (0,+\infty)$$

Una propiedad muy importante de la F Fisher es la siguiente:

$$F_{\nu_1;\nu_2;\alpha} = \frac{1}{F_{\nu_2;\nu_1;1-\alpha}}$$

Partimos del estudio de dos variables aleatorias independientes a las que llamaremos X e Y. Ambas variables deben tener obligatoriamente naturaleza normal.

Extraemos sendas muestras de ambas poblaciones de tamaño n_x y n_y (no tienen obligatoriamente que adoptar el mismo valor).

$$X \to N(\mu_x;\sigma^2{}_x) \qquad Y \to N(\mu_y;\sigma^2{}_y)$$

$$\frac{(n_x-1)\,S^2{}_x}{\sigma_x^2} = \chi^2{}_{n_x-1} \qquad \frac{(n_y-1)\,S^2{}_y}{\sigma_y^2} = \chi^2{}_{n_y-1}$$

$$\frac{\dfrac{\chi^2{}_{n_x-1}}{n_x-1}}{\dfrac{\chi^2{}_{n_y-1}}{n_y-1}} = F_{n_x-1;n_y-1} \Rightarrow$$

$$\Rightarrow \frac{\dfrac{\dfrac{(n_x-1)\,S^2{}_x}{\sigma_x^2}}{n_x-1}}{\dfrac{\dfrac{(n_y-1)\,S^2{}_y}{\sigma_y^2}}{n_y-1}} = \frac{\dfrac{S^2{}_x}{\sigma_x^2}}{\dfrac{S^2{}_y}{\sigma_y^2}} = F_{n_x-1;n_y-1}$$

Esta expresión la vamos a emplear para saber si las varianzas de las dos variables iniciales de partida son iguales. Esta propiedad de igualdad de varianzas recibe el nombre de **homocedasticidad.**

En el caso de que ambas poblaciones tengan la misma varianza, esto es, en el caso de que las poblaciones sean homocedásticas, la expresión anterior se transforma en:

$$\frac{\frac{S^2_x}{\sigma_x^2}}{\frac{S^2_y}{\sigma_y^2}} = F_{n_x-1;n_y-1}$$

$$Si\ \sigma_x^2 = \sigma_y^2 \Rightarrow \frac{S^2_x}{S^2_y} = F_{n_x-1;n_y-1}$$

V.E. Ejemplo.

Partimos de los datos de la ciudad X que han aparecido en un ejemplo anterior. El mismo estudio se lleva a cabo en otra ciudad extrayendo una muestra de 11 viviendas con los siguientes precios de alquiler: 96; 96; 93; 95; 97; 103; 96; 101; 97; 96; 99. Nos piden determinan si el precio del alquiler en ambas ciudades es homocedástico.

X: Precio del alquiler de 1 vivienda en la ciudad X.

Y: Precio del alquiler de 1 vivienda en la ciudad Y.

Nat(X) --> Debemos suponerla NORMAL (el modelo que hemos planteado exige que las variables de partida tengan naturaleza normal).

Nat Y --> Debemos considerarla NORMAl por el mismo motivo.

$\mu_x = ?$ $\sigma^2_x = ?$ $\mu_y = ?$ $\sigma^2_y = ?$

A continuación extraemos de las poblaciones sendas muestras de tamaños 7 y 11 respectivamente, con lo que tendremos por un lado 7 VAIID (a pesar de tratarse de muestreo sin reemplazamiento la población es infinita) y por el otro otras 11 VAIID.

Definimos dos nuevas variables que serán "Varianza en el precio del alquiler en una muestra de tamaño 7 extraída de la ciudad X" (S^2_7) y "Varianza en el precio del alquiler en una muestra de tamaño 11 extraída de la ciudad Y" (S^2_{11}):

$$S^2_7 = \frac{\sum_{i=1}^{7} (X_i - \overline{X})^2}{6} \qquad S^2_{11} = \frac{\sum_{i=1}^{11} (Y_i - \overline{Y})^2}{10}$$

La cuestión que se nos plantea es conocer si ambas poblaciones son homocedásticas, esto es, si sus varianzas poblacionales son iguales. Para ello deberemos seguir los siguientes pasos:

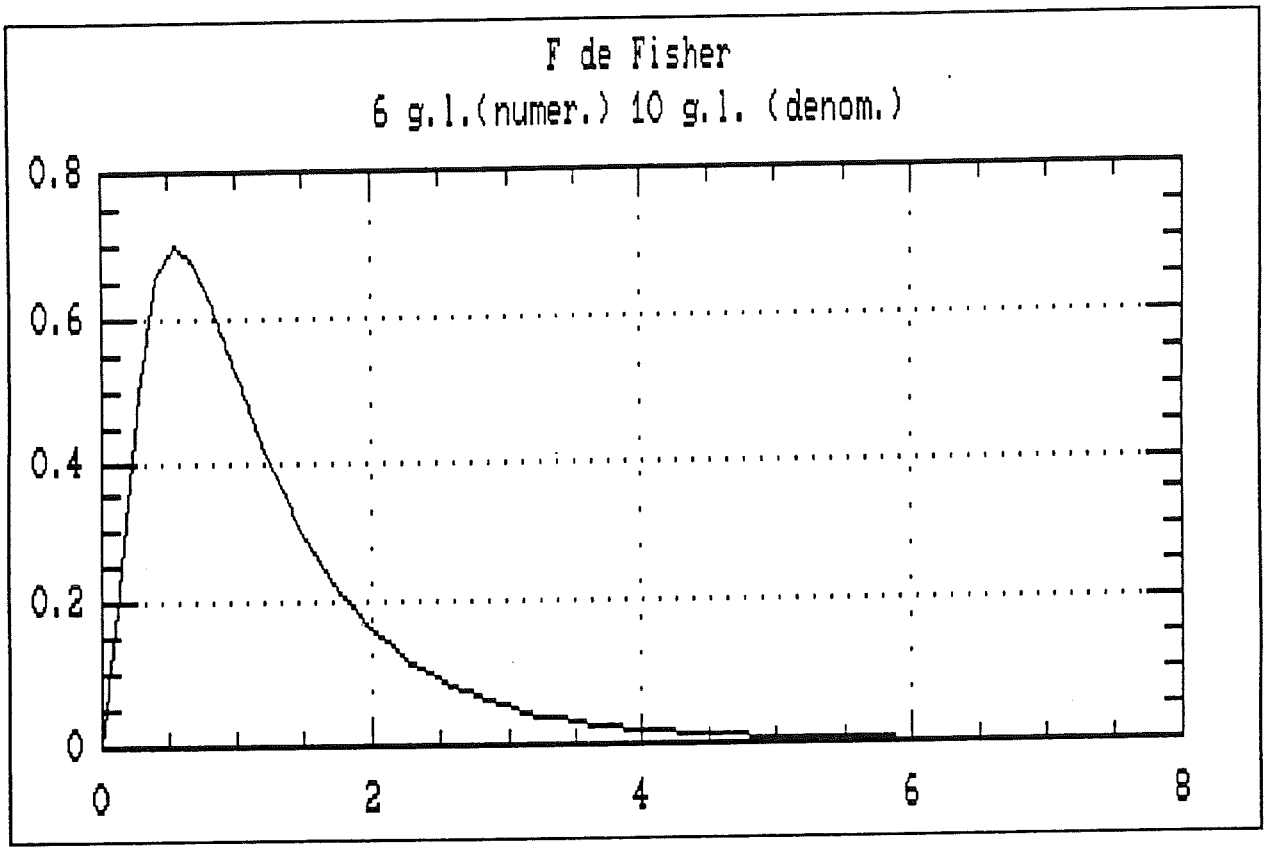
1º) Calcular la puntuación de la F Fisher con las varianzas obtenidas en las muestras:

$$S^2_{11} = 2{,}822^2 \quad S^2_7 = 2{,}828^2$$

$$S^2_7 / S^2_{11} = 1{,}005$$

Esta puntuación corresponde a una F de Fisher de 6 ($n_x - 1$) grados de libertad en el numerador y 10 ($n_y - 1$) grados de libertad en el denominador.

2º) Representar la $F_{6,10}$ y determinar arbitrariamente un área central o nivel de confianza. En este caso, elegiremos el 95%



Consultando las oportunas tablas, los valores que encierran un área central del 95% son 0,18 y 4,07. Dado que la puntuación F Fisher (1,005) obtenida a partir de los datos de la muestra se encuentra dentro del área central, se confirma el supuesto de partida de que ambas poblaciones son homocedásticas. Esto es, las diferencias observadas entre las varianzas de ambas muestras se deben al azar y no resultan ser significativas para un nivel del 5%

Si la puntuación F Fisher hubiese caído fuera del área central subjetivamente determinada, rechazaríamos el supuesto de partida y diríamos que las poblaciones son heterocedásticas (sus varianzas no son iguales). Las diferencias entre las varianzas muestrales resultarían ser entonces significativas.

IV.D. Distribución muestral de la media

Análisis que involucran a dos poblaciones

En ocasiones puede surgir la necesidad de comparar las medias de dos poblaciones distintas. En estos casos, vamos a distinguir cuatro posibles situaciones, en función de cómo sean las varianzas de las dos poblaciones objeto de estudio:

- Las varianzas poblacionales son **conocidas y distintas.**
- Las varianzas poblacionales son **conocidas e iguales.**
- Las varianzas poblacionales son **desconocidas pero iguales.**
- Las varianzas poblacionales son **desconocidas y diferentes.**

En todo caso partiremos de dos variables aleatorias X e Y que deberán ser **independientes y normales.** Tendremos así:

$$X \approx N(\mu_x; \sigma_x^2) \qquad Y \approx N(\mu_y; \sigma_y^2)$$

Sabemos entonces que:

$$\overline{X} \approx N(\mu_x; \frac{\sigma_x^2}{n_x}) \qquad \overline{Y} \approx N(\mu_y; \frac{\sigma_y^2}{n_y})$$

POR TANTO:

$$\overline{X}-\overline{Y} \Rightarrow \textit{Naturaleza normal}$$

$$\mu_{\overline{X}-\overline{Y}} = \mu_x - \mu_y$$

$$\sigma^2_{\overline{X}-\overline{Y}} = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$$

$$\overline{X}-\overline{Y} \approx N(\mu_x-\mu_y; \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y})$$

* En el primer caso (varianzas poblacionales conocidas pero distintas), tipificando la distribución muestral de la diferencia de medias, llegaremos a la siguiente expresión:

$$\frac{\overline{X}-\overline{Y} - (\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} = Z$$

* En el segundo caso (varianzas poblacionales conocidas e iguales) la expresión será:

$$\frac{\overline{X}-\overline{Y} - (\mu_x-\mu_y)}{\sqrt{\frac{\sigma^2}{n_x}+\frac{\sigma^2}{n_y}}} = Z$$

donde σ^2 es la varianza común

Ejemplo: Se toman dos muestras aleatorias independientes de 64 observaciones cada una de dos poblaciones normales con las mismas medias y con desviaciones estándar $\sigma_x = 6,4$ y $\sigma_y = 7,2$ respectivamente. Calcular la probabilidad de que la diferencia en valor absoluto entre las dos medias muestrales exceda de 0,6.

X: Valor de 1 observación en la población X

Y: Valor de 1 observación en la población Y

Nat X --> NORMAL $\mu_x = ?$ $\sigma^2_x = 6,4^2$
Nat Y --> NORMAL $\mu_y = ?$ $\sigma^2_y = 7,2^2$

De la población infinita X extraemos (MSR) una muestra de 64 observaciones, con lo que obtendremos 64 VAIID. Idénticamente, de la población infinita Y extraeremos (MSR) una muestra 64 observaciones y tendremos también otras 64 VAIID.

Definiremos las dos distribuciones muestrales como: "Media del valor obtenido en una muestra de tamaño 64 de la población X" ($\overline{X}_{\infty}$) y "Media del valor obtenido en una muestra de tamaño 64 de la población Y" ($\overline{Y}_{\infty}$).

$$\overline{X}_{64} = \frac{\sum_{i=1}^{64} X_i}{64} \qquad \overline{Y}_{64} = \frac{\sum_{i=1}^{64} Y_i}{64}$$

$$\overline{X}_{64} \approx NORMAL \qquad \overline{Y}_{64} \approx NORMAL$$

$$\mu_{\overline{X}_{64}} = \mu_x \qquad \mu_{\overline{Y}_{64}} = \mu_y$$

$$\sigma^2_{\overline{X}_{64}} = \frac{\sigma_x^2}{n_x} = \frac{6.4^2}{64} \qquad \sigma^2_{\overline{Y}_{64}} = \frac{\sigma_y^2}{n_y} = \frac{7.2^2}{64}$$

$$\overline{X}_{64} - \overline{Y}_{64} \approx NORMAL$$

$$\mu_{\overline{X}_{64}-\overline{Y}_{64}} = \mu_x - \mu_y = 0 \ (son\ iguales)$$

$$\sigma^2_{\overline{X}_{64}-\overline{Y}_{64}} = \frac{6.4^2}{64} + \frac{7.2^2}{64} = 1.45$$

La probabilidad que nos piden calcular puede expresarse así:

$$P[|\overline{X}_{64}-\overline{Y}_{64}| > 0.6] = 1 - P[|\overline{X}_{64}-\overline{Y}_{64}| < 0.6]$$

$$= 1 - P[-0.6 < \overline{X}_{64}-\overline{Y}_{64} < +0.6] =$$

$$= 1 - P[-0.498 < Z < 0.498] \approx 0.617$$

La solución obtenida puede interpretarse del siguiente modo: aproximadamente en 62 de cada 100 ocasiones en que analicemos muestras de 64 observaciones en cada población, la diferencia de ambas medias muestrales excederá en valor absoluto a 0,6.

* Para el tercer caso (varianzas desconocidas, pero **por la prueba de homocedasticidad** tenemos evidencia de que no son significativamente diferentes) recordemos que con varianzas conocidas e iguales hemos visto que:

$$\frac{\overline{X}-\overline{Y}-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma^2}{n_x}+\frac{\sigma^2}{n_y}}} = Z$$

Sabemos además que:

$$\frac{(n_x-1)S_x^2}{\sigma^2} = \chi^2_{n_x-1} \qquad \frac{(n_y-1)S_y^2}{\sigma^2} = \chi^2_{n_y-1}$$

$$W = \frac{(n_x-1)S_x^2}{\sigma^2} + \frac{(n_y-1)S_y^2}{\sigma^2} = \chi^2_{n_x+n_y-2}$$

Utilizando las expresiones anteriores vamos a construir una expresión t Student del siguiente modo:

$$t_{n_x+n_y-2} = \frac{Z}{\sqrt{\frac{W}{n_x+n_y-2}}} =$$

$$= \frac{\frac{\overline{X}-\overline{Y}-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma^2}{n_x}+\frac{\sigma^2}{n_y}}}}{\sqrt{\frac{\frac{(n_x-1)S_x^2}{\sigma^2}+\frac{(n_y-1)S_y^2}{\sigma^2}}{n_x+n_y-2}}} =$$

$$t_{n_x+n_y-2} = \frac{\overline{X}-\overline{Y} - (\mu_x-\mu_y)}{\sqrt{\frac{S^2{}_p}{n_x} + \frac{S^2{}_p}{n_y}}}$$

Donde $S^2{}_p$ recibe el nombre de varianza combinada (la prueba se denomina "pooled variance t test"). La expresión de $S^2{}_p$ es:

$$S_p^2 = \frac{(n_x-1)\,S^2{}_x + (n_y-1)\,S^2{}_y}{n_x+n_y-2}$$

Ejemplo: En las dos ciudades X e Y de ejemplos anteriores, en las que medíamos el precio del alquiler de las viviendas, se pide comprobar si el precio medio del alquiler es el mismo para ambas ciudades.

X: Precio del alquiler de 1 vivienda en la ciudad X.
Y: Precio del alquiler de 1 vivienda en la ciudad Y.

Nat (X) --> suponemos NORMAL $\mu_x = ?$ $\sigma^2{}_x = ?$

Nat (Y) --> suponemos NORMAL $\mu_y = ?$ $\sigma^2{}_y = ?$

Tomamos una muestra de 7 viviendas de la ciudad X, con lo que dispondremos de 7 VAIID y otra muestra de 11 viviendas de la ciudad Y que serán otras 11 VAIID.

Definimos dos nuevas variables:

$$\overline{X}_7 = \frac{\sum_{i=1}^{7} X_i}{7} \qquad \overline{Y}_{11} = \frac{\sum_{i=1}^{11} Y_i}{11}$$

$$\mu_{\overline{X}_7} = \mu_x = ? \qquad \mu_{\overline{Y}_{11}} = \mu_y = ?$$

$$\sigma^2_{\overline{X}_7} = \frac{\sigma^2{}_x}{7} = ? \qquad \sigma^2_{\overline{Y}_{11}} = \frac{\sigma^2{}_y}{11} = ?$$

En cada una de las muestras extraídas de cada población calculamos la media muestral y la varianza muestral, obteniendo los siguientes resultados:

$$\overline{X}_7 = 102 \qquad S^2{}_7 = 2.828^2$$

$$\overline{Y}_{11} = 97.18 \qquad S^2{}_{11} = 2.822^2$$

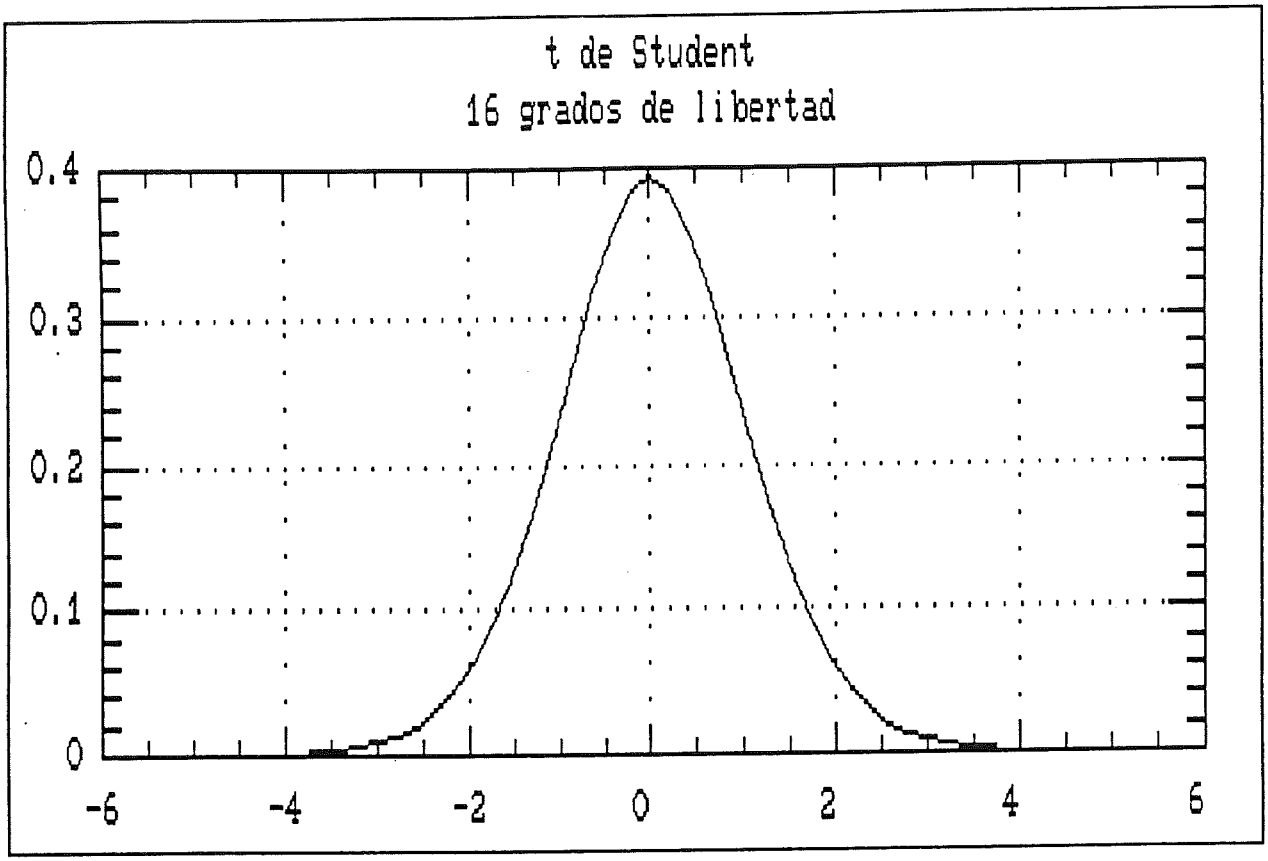
Ya que nos encontramos en el caso de varianzas desconocidas pero iguales (recordemos que en un ejercicio anterior no hemos encontrado pruebas de que las varianzas de ambas poblaciones fueran distintas), debemos calcular la varianza ponderada $S^2{}_p$ y obtener la puntuación t Student de 7+11-2=16 grados de libertad **partiendo del supuesto de que ambas poblaciones tienen medias iguales:**

$$S^2{}_p = \frac{6 \; x \; 2.828^2 + 10 \; x \; 2.822^2}{16} = 2.824^2$$

$$\frac{(\overline{X}_7 - \overline{Y}_{11}) - (\mu_x - \mu_y)}{\sqrt{\frac{S^2{}_p}{n_x} + \frac{S^2{}_p}{n_y}}} = t_{16}$$

$$\frac{(102 - 97.18) - 0}{\sqrt{\frac{2.824^2}{7} + \frac{2.824^2}{11}}} = 3.5301 \rightarrow t_{16}$$

Sólo resta representar la t Student de 16 grados de libertad, seleccionar un área central arbitraria (en este caso del 95%) y observar dónde queda situada la puntuación que hemos obtenido con la información de las muestras. En función de donde quede la puntuación obtenida, dentro o fuera de la zona central, aceptaremos o rechazaremos el supuesto de partida de que ambas poblaciones tienen medias iguales.



Los valores entre los que se encuentra encerrada un área correspondiente al 95% (subjetivamente decidido) son -2,12 y +2,12. El valor de la t Student obtenido en la prueba (+3,5301) se encuentra claramente fuera de esta zona central, con lo que no se mantiene el supuesto de partida de que las medias de ambas poblaciones son iguales.

No podemos afirmar que el precio medio del alquiler en ambas ciudades sea el mismo. Es más, tenemos pruebas para sospechar que el precio medio del alquiler en la ciudad X es superior al precio medio del alquiler en la ciudad Y.

Sin embargo, al realizar esta afirmación (es decir, al rechazar el supuesto de partida) podemos equivocarnos. Existe una probabilidad del 5% (nivel de significación) de que afirmemos que el precio medio del alquiler no es el mismo en ambas ciudades y nos estemos equivocando porque realmente sean iguales. Es decir, existe una probabilidad del 5% de que, a la luz de la información de ambas muestras, rechacemos el supuesto de partida y éste sea realmente cierto.

* Debemos considerar por último qué sucede cuando las varianzas poblacionales son desconocidas y la prueba de homocedasticidad nos indica que, además, no son iguales.

En este caso, poco frecuente, denominado "separate variance t test", se utiliza la denominada aproximación de Welch, que se expresa como sigue:

$$\frac{(\overline{X}-\overline{Y}) - (\mu_x-\mu_y)}{\sqrt{\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}}} = t$$

con grados de libertad:

$$\frac{(n_x-1)(n_y-1)}{(n_y-1)c^2+(1-c)^2(n_x-1)}$$

$$donde\ c = \frac{\frac{s^2{}_x}{n_x}}{\frac{s^2{}_x}{n_x}+\frac{s^2{}_y}{n_y}}$$

En la expresión anterior, si los grados de libertad resultan no ser un número entero, se suelen redondear **por defecto** al entero más próximo antes de consultar las tablas de la t Student.

En la práctica, si los grados de libertad de la t resultante son un número suficientemente grande, se opta por aproximar la t a una distribución Z.

VI. DISTRIBUCIONES MUESTRALES DERIVADAS DE LA BINOMIAL

La variable que estudiaremos en la población de partida no va a ser -como hasta ahora- de naturaleza cuantitativa, sino cualitativa. De entre éstas vamos a centrarnos en el estudio de variables cualitativas dicotómicas o variables de BOOLE. Estas variables, que representaremos con la letra B, tienen tan sólo dos modalidades de aparición, a las que técnicamente denominaremos EXITO y FRACASO.

VI.A. La distribución muestral del número de éxitos

De la población de partida conocemos un parámetro, que designaremos por P, y que representa el porcentaje o la probabilidad de aparición de la modalidad éxito en la población. En consecuencia conocemos también Q, que es la probabilidad o porcentaje poblacional de fracasos, y que es igual a 1-P. Supongamos que de esta población de partida (de tamaño N) extraemos todas (las K) muestras posibles de tamaño n. En cada una de ellas calcularemos un estadístico, que designaremos por "r", y que definiremos como el "número de éxitos en una muestra de tamaño n". "r" podrá tomar n+1 valores que irán desde 0 (ningún éxito en la muestra) hasta n (toda la muestra formada por éxitos).

Evidentemente r es una variable y como tal puede variar de valor de muestra a muestra. Los k valores de r obtenidos en las k posibles muestras de tamaño n que se pueden extraer de esa población constituyen LA DISTRIBUCION MUESTRAL DE EXITOS. ¿Cuáles son entonces los parámetros -media y varianza- y la naturaleza de la variable "r"?. Debemos considerar dos casos:

a) En el caso de MSR la naturaleza de la distribución muestral del número de éxitos (r) es **hipergeométrica**, y sus parámetros valor esperado y varianza son:

$$E[r] = nP \qquad Var(r) = nPQ\left(\frac{N-n}{N-1}\right)$$

$$Nat(r) \rightarrow Hipergeométrica$$

$$Discreta: puede\ tomar\ (n+1)\ valores$$

La probabilidad de que en una muestra de tamaño n extraída de una población de tamaño N, r tome un determinado valor k viene dada por la siguiente expresión del modelo hipergeométrico:

$$P[r=k] = \frac{\binom{NP}{k}\binom{NQ}{n-k}}{\binom{N}{n}}$$

Ejemplo: En una población de 12 personas un tercio de ellas están enfermas. Seleccionada sin reemplazamiento una muestra de 3 personas, calcular la probabilidad de que en dicha muestra haya ninguna, una, dos y tres personas enfermas.

B: Una persona de la población está enferma (éxito; P=1/3).
no está enferma (fracaso; Q=2/3).

El tamaño de la población es 12 (N=12). De esta población extraemos sin reemplazamiento una muestra de tres personas y definimos una variable a la que designaremos por r y que expresará el **"Número de personas enfermas en una muestra de 3 personas"**. Evidentemente r podrá tomar 4 valores que irán de 0 a 3. La naturaleza de r es hipergeométrica y sus parámetros son:

$$E[r] = nP = 3 \times \frac{1}{3} = 1$$

$$Var(r) = nPQ\left(\frac{N-n}{N-1}\right) = 3 \times \frac{1}{3} \times \left(\frac{12-3}{12-1}\right) = 0.545$$

Las probabilidades que nos piden son entonces:

$$P[r=0] = \frac{\binom{4}{0}\binom{8}{3}}{\binom{12}{3}} = 0.2545 \qquad P[r=1] = \frac{\binom{4}{1}\binom{8}{2}}{\binom{12}{3}} = 0.5091$$

$$P[r=2] = \frac{\binom{4}{2}\binom{8}{1}}{\binom{12}{3}} = 0.2182 \qquad P[r=3] = \frac{\binom{4}{3}\binom{8}{0}}{\binom{12}{3}} = 0.0182$$

Interpretando el resultado diríamos que, aproximadamente, de cada 100 veces en que analicemos muestras de 3 personas extraídas sin reemplazamiento de una población de 12 personas en la cual la tercera parte están enfermas, en 25 ocasiones encontraremos que no hay ninguna persona enferma, en 51 ocasiones encontraremos una persona enferma, en 22 ocasiones encontraremos dos personas enfermas y en 2 ocasiones encontraremos 3 personas enfermas.

En este ejemplo, la utilización de la distribución hipergeométrica ha sido relativamente sencilla. Sin embargo, dado que la distribución hipergeométrica lleva aparejada la utilización de números combinatorios, los cálculos se van haciendo más complejos a medida que el tamaño de la población y/o el tamaño de la muestra van aumentando. **Cuando los valores de N y/o n que estamos manejando sean grandes, el modelo hipergeométrico se puede aproximar al modelo binomial.**

Ejemplo: Supongamos que de una población de 1 millón de consumidores se extrae al azar una muestra sin reeemplazamiento de 16 de ellos para hacerles probar un nuevo producto. Por experiencias pasadas, se sabe que ese tipo de productos gusta al 30% de los consumidores. Calcular la probabilidad de que como mucho a seis consumidores de la muestra les guste el producto.

B: A un consumidor le gusta el producto (éxito; P=0,3)
no le gusta el producto (fracaso; Q=0,7).

El tamaño de la población es 1.000.000 y de ella extraemos mediante muestreo sin reemplazamiento una muestra de 16 consumidores, definiendo así una nueva variable que designaremos por r, y que expresará el **"Número de personas a las que gusta el producto en una muestra de 16 consumidores"**. La variable puede tomar 17 valores, que podrán variar desde 0 hasta 16. Su naturaleza es hipergeométrica y sus parámetros son:

$$E[r] = nP = 16 \times 0.3 = 4.8$$

$$Var(r) = nPQ\left(\frac{N-n}{N-1}\right) =$$

$$= 16 \times 0.3 \times 0.7 \left(\frac{1000000-16}{1000000-1}\right) \approx 3.36$$

La probabilidad que nos piden viene dada entonces por la siguiente expresión:

$$P[r \leq 6] = \sum_{k=0}^{6} \frac{\binom{1000000 \times 0.3}{k} \binom{1000000 \times 0.7}{16-k}}{\binom{1000000}{16}}$$

El cálculo de esta probabilidad a través del modelo hipergeométrico resulta, si no imposible, sí sumamente engorroso. Sin embargo hay que señalar que, bien resulte imposible, bien lo descartemos por su dificultad, el modelo hipergeométrico es **el único que ofrece el resultado exacto para la distribución muestral del número de éxitos en el caso de muestreo sin reemplazamiento.**

Ahora bien, como ya hemos indicado, el modelo binomial puede ofrecer, para valores de N lo suficientemente grandes, una buena aproximación al modelo hipergeométrico. Según el modelo binomial podemos expresar la probabilidad pedida del siguiente modo:

$$P[r \leq 6] = \sum_{k=0}^{6} \binom{16}{k} 0.3^k \, 0.7^{16-k} = 0.8246$$

Interpretando el resultado diríamos que aproximadamente 82 de cada 100 veces en que analice muestras de 16 consumidores, encontraré que como mucho a 6 de ellos les gusta el nuevo producto.

En ocasiones puede interesarnos calcular la probabilidad de un rango muy amplio de valores de r. En casos como éste la utilización del modelo binomial resulta también bastante engorrosa (limitación de tablas...), por lo que sería deseable poder aproximar la distribución binomial (que es una distribución discreta) a una distribución continua para facilitar el cálculo de probabilidades. Vamos a ver entonces como, bajo determinadas condiciones, la distribución binomial puede ser aproximada a una distribución normal.

Ejemplo: Se sabe que en una comunidad el 55% de la población apoya a un determinado candidato político. Encuestada una muestra aleatoria de 900 votantes, calcular la probabilidad de que más de 540 de ellos voten a favor del citado candidato.

B: Un votante apoya al candidato (éxito ; P=0,55).
no apoya al candidato (fracaso ; Q=0,45).

El tamaño de la población es desconocido. Mediante muestreo sin reemplazamiento (ya que no queremos preguntar dos veces a la misma persona) extraemos de esa población una muestra de 900 votantes, definiendo una nueva variable (distribución muestral del número de éxitos) que designaremos por r y que expresará el "Número de votantes que apoyan al candidato en una muestra de 900 votantes". La variable r podrá tomar 901 valores que podrán encontrarse entre 0 y 900. La naturaleza de r es hipergeométrica, y sus parámetros son:

$$E[r] = nP = 900 \; x \; 0.55 = 495$$

$$Var(r) = nPQ\left(\frac{N-n}{N-1}\right) = ?$$

La probabilidad pedida, expresada a través del modelo hipergeométrico, viene dada por:

$$P[r>540] = \sum_{k=541}^{900} \frac{\binom{0.55N}{k}\binom{0.45N}{900-k}}{\binom{N}{900}}$$

Este es el único modelo que proporciona el resultado exacto a la probabilidad pedida. Sin embargo, en este caso el cálculo de la probabilidad pedida según el modelo hipergeométrico no es ya engorroso sino imposible, ya que desconocemos el tamaño de la población. Dado que el tamaño de la población puede considerarse suficientemente grande, podemos aproximar el modelo hipergeométrico al modelo binomial, recordando que el resultado que en todo caso obtengamos no será exacto sino tan sólo una aproximación al verdadero valor que se obtendría empleando (si fuera posible) el modelo hipergeométrico. La probabilidad pedida, aproximada por el modelo binomial, viene dada por:

$$P[r>540] = \sum_{k=541}^{900} \binom{900}{k} 0.55^{k} \; 0.45^{900-k}$$

Este calculo a través del modelo binomial resulta aunque no imposible, sí sumamente dificultoso. Estaríamos obligados bien a consultar en las tablas todos las probabilidades para los valores que están comprendidos entre el 541 y el 900, bien a calcular mediante números combinatorios todas y cada una de esas probabilidades y sumarlas. Además de esta dificultad, el resultado que obtendríamos tampoco sería exacto, con lo que no se justifica un esfuerzo de esas características..

Afortunadamente podemos aproximar la distribución binomial (discreta) a una distribución normal (continua), lo que facilita los cálculos. Para poder llevar a cabo esta aproximación debe cumplirse que **el producto nPQ sea al menos igual a 3**, y cuanto mayor sea el valor de nPQ de mejor calidad será el ajuste del modelo binomial al modelo normal. En nuestro ejemplo tenemos:

$$nPQ \geq 3$$

$$900 \times 0.55 \times 0.45 = 222.75 \geq 3$$

Como el valor obtenido en nPQ supera ampliamente a 3, podemos suponer que el ajuste del modelo binomial al modelo normal será aceptablemente bueno. Tendremos así:

$$r \rightarrow X$$

$$Discreta \rightarrow Continua$$

$$Hipergeométrica \rightarrow Binomial \rightarrow Normal$$

$$E[X] = E[r] = nP = 900 \times 0.55 = 495$$

$$Var(X) = Var(r) = nPQ = 900 \times 0.55 \times 0.45 = 222.75$$

De este modo podemos calcular fácilmente la probabilidad pedida. Sin embargo, por el hecho de haber ajustado una distribución de probabilidad discreta (r) a una distribución de probabilidad continua (X), debemos aplicar lo que se conoce como **corrección de continuidad**, y que consiste en la suma o resta (según como sea la pregunta probabilística formulada) de media unidad de medida a los valores de la distribución de probabilidad continua para acomodarlos a la distribución de probabilidad discreta.

Al tratar de calcular la probabilidad de que el número de votantes a favor del candidato sea estrictamente superior a 540, ¿desde dónde debemos empezar a calcular en el caso de una distribución continua?. En el caso de que la distribución sea discreta es muy claro que el valor superior a 540 es 541. En el caso de que la distribución de probabilidad sea continua, se emplea la corrección por continuidad que consiste, en este ejemplo, en añadir al valor media unidad de medida:

$$P[r>540] \approx P[X>540.5] = 1-P[X\leq 540.5] =$$

$$= 1-P[z\leq \frac{540.5-495}{\sqrt{222.75}}] =$$

$$= 1-P[z\leq 3.048] = 0.0011$$

Podemos interpretar este resultado del siguiente modo: aproximadamente 1 de cada 1000 veces en que analice mediante MSR muestras de 900 votantes procedentes de esa comunidad, obtendré que más de 540 de ellos apoyan al candidato.

b) Recordemos que queda pendiente el análisis cuando se **muestrea con reemplazamiento** la población de partida. En este caso, la naturaleza de la distribución muestral del número de éxitos será **binomial**, siendo también de aplicación la aproximación binomial a la normal cuando nPQ sea al menos igual a 3.

$$E[r] = nP \qquad Var(r) = nPQ$$

$$Nat(r) \rightarrow Binomial$$

$$Discreta: puede\ tomar\ (n+1)\ valores$$

$$P[r=k] = \binom{n}{k} P^k Q^{n-k}$$

$$Si\ nPQ\geq 3\ entonces:$$

$$r \rightarrow X$$

$$Discreta \rightarrow Continua$$

$$Binomial \rightarrow Normal$$

VI.B. Distribución muestral de la proporción de éxitos

El estadístico que calcularemos en las muestras de tamaño n será $\overline{p}$ y expresará la **"proporción de éxitos en una muestra de tamaño n"**. Esta nueva distribución se obtiene a partir de la distribución muestral del número de éxitos del siguiente modo:

$$\overline{p} = \frac{r}{n} = \frac{Número\ de\ éxitos}{Tamaño\ de\ muestra}$$

$$E(\overline{p}) = \frac{1}{n} E(r) = \frac{1}{n} nP = P$$

$$Var(\overline{p}) = Var(\frac{r}{n}) = \frac{1}{n^2} Var(r) =$$

$$= \frac{1}{n^2} nPQ \left(\frac{N-n}{N-1}\right) = \frac{PQ}{n} \left(\frac{N-n}{N-1}\right)$$

$$Nat(\overline{p}) = Nat(r)$$

Recordando, en casos de MSR la distribución muestral $\overline{p}$ será de naturaleza hipergeométrica. Si el tamaño de muestra y/o el de la población son grandes, se podrá aproximar por el modelo binomial. Si, por último, se cumple que el producto nPQ vale al menos 3 se podrá aproximar al modelo continuo normal usando la corrección de continuidad. En casos de MCR, la distribución muestral $\overline{p}$ será de naturaleza binomial. Si se cumple que el producto nPQ vale al menos 3 se podrá aproximar al modelo continuo normal usando la corrección de continuidad.

Ejemplo: Se extrae una muestra de 900 votantes de una comunidad en la que el 55% apoya a un determinado candidato político. Nos piden obtener la probabilidad de que más de un 60% de los votantes de la muestra apoyen al candidato.

B:Un votante de la comunidad apoya al candidato (éxito ; P=0.55)
no apoya al candidato (fracaso ; Q=0.45)

El tamaño (N) de la población es desconocido. De esta población extraemos mediante MSR una muestra aleatoria de 900 votantes. Sobre esta muestra definimos una nueva variable que

designaremos por $\bar{p}$ y que expresará la "**proporción de votantes que apoyan al candidato en una muestra de 900**".

$\bar{p}$ es una variable discreta que puede tomar 901 valores, que se encontrarán entre 0/900 (ningún éxito en la muestra = proporción 0 de éxitos) y 900/900 (todo éxitos en la muestra = proporción 1 de éxitos). Sus parámetros y naturaleza serán:

$$E(\bar{p}) = P = 0.55$$

$$Var(\bar{p}) = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right) =$$

$$= \frac{0.55 \; x \; 0.45}{900}\left(\frac{N-900}{N-1}\right) = ?$$

$$Nat(\bar{p}) \rightarrow Hipergeométrica$$

La probabilidad pedida vendrá dada de forma exacta por:

$$P[\bar{p} > 0.6] = P[\bar{p} > \frac{540}{900}] = P[r > 540] =$$

$$= \sum_{k=541}^{900} \frac{\binom{0.55\,N}{k}\binom{0.45\,N}{900-k}}{\binom{N}{900}} = ?$$

Dado que no conocemos el tamaño N de la población, el cálculo de la probabilidad pedida a través del modelo hipergeométrico resulta inviable (y si se conociese N, el cálculo resultaría sumamente tedioso).

Sin embargo, y al igual que hicimos en el caso de la distribución muestral del número de éxitos, resulta posible para valores grandes de N y/o de n la aproximación de la distribución hipergeométrica a la binomial, por lo que podremos aproximar el resultado del siguiente modo:

$$P[\overline{p} > 0.6] = P[r > 540] =$$

$$= \sum_{k=541}^{900} \binom{900}{k} 0.55^k \, 0.45^{900-k}$$

Con todo, el cálculo de esta probabilidad sigue siendo sumamente engorroso, por lo que trataremos de aproximar la distribución binomial a la distribución normal, ya que:

$$nPQ \geq 3$$

$$900 \cdot 0.55 \cdot 0.45 = 222.75 \geq 3$$

También en este caso tendremos que utilizar la corrección por continuidad, que en el caso de la distribución muestral de la proporción de éxitos consistirá en sumar o restar (según cómo se haya formulado la pregunta probabilística) al valor de la variable continua 1/2n.

$$\overline{p} \rightarrow X$$

$$Discreta \rightarrow Continua$$

$$Binomial \rightarrow Normal$$

$$E(X) = E(\overline{p}) = 0.55$$

$$Var(X) = Var(\overline{p}) = \frac{PQ}{n} = \frac{0.55 \times 0.45}{900} = 0.000275$$

De este modo, la probabilidad pedida, expresada a través de la aproximación a la distribución normal vendrá dada por:

$$P[\overline{p} > 0.6] = P[X > 0.6 + \frac{1}{2 \cdot 900}] =$$

$$= 1 - P[X < 0.6 + \frac{1}{2 \cdot 900}] = 1 - P[Z < 3.049] = 0.0011$$

Como se puede comprobar, el resultado es el mismo que en el ejemplo anterior. **Hemos resuelto el mismo problema de dos formas distintas. En primer lugar a través de la distribución muestral del número de éxitos y en segundo lugar a través de la distribución muestral de la proporción de éxitos.** Toda pregunta sobre la distribución muestral de la proporción de éxitos podrá resolverse a través de la distribución muestral del número de éxitos y viceversa.

Por último es preciso señalar que en algunos casos, en vez de utilizar la aproximación normal a la binomial, es mejor utilizar otras aproximaciones. Por ejemplo, si

$$n \to \infty$$

$$P \to 0$$

la distribución binomial se ajusta mejor a una distribución de Poisson con

$$\lambda = nP$$

VI.C. Distribuciones derivadas de la binomial.

Análisis que involucran a dos poblaciones

Vamos a estudiar el comportamiento tanto de la distribución muestral del número de éxitos como de la distribución muestral de la proporción de éxitos cuando se estudian en dos muestras aleatorias extraídas de dos poblaciones independientes X e Y.

Ejemplo: Los resultados de una investigación de mercados demostraron que tanto en un país X como en otro país Y un 65% de los consumidores estarían dispuestos a adquirir un nuevo producto. Hallar la probabilidad de que en dos muestras aleatorias obtenidas en cada país, compuesta cada una por 200 consumidores, haya

a) una diferencia superior a 10 consumidores;
b) una diferencia superior al 10% en la proporción de consumidores;

que estarían dispuestos a adquirir el nuevo producto.

B_x : Un consumidor del país X estaría (éxito; $P_x = 0.65$)
no estaría (fracaso; $Q_x = 0.35$)
dispuesto a comprar el nuevo producto.

B_y : Un consumidor del país Y estaría (éxito; $P_y = 0.65$)
no estaría (fracaso; $Q_y = 0.35$)
dispuesto a comprar el nuevo producto.

Debemos suponer además que las poblaciones de ambos países son independientes.

En cada uno de los países seleccionamos mediante MSR a 200 personas y definimos dos nuevas variables aleatorias:

r_x: "Número de consumidores dispuestos a adquirir el nuevo producto en una muestra de 200 personas seleccionada en el país X".

r_y: "Número de consumidores dispuestos a adquirir el nuevo producto en una muestra de 200 personas seleccionada en el país Y".

La naturaleza de las dos nuevas variables que hemos definido es hipergeométrica y sus parámetros son:

$$E[r_x] = n_x P_x = 200 \cdot 0.65 = 130$$

$$E[r_y] = n_y P_y = 200 \cdot 0.65 = 130$$

$$Var(r_x) = n_x P_x Q_x \left(\frac{N_x - n_x}{N_x - 1}\right) \approx$$

$$\approx 200 \cdot 0.65 \cdot 0.35 \cdot 1 = 45.5$$

$$Var(r_y) = n_y P_y Q_y \left(\frac{N_y - n_y}{N_y - 1}\right) \approx$$

$$\approx 200 \cdot 0.65 \cdot 0.35 \cdot 1 = 45.5$$

A continuación ajustamos directamente las dos variables definidas a dos nuevas variables normales que designaremos por X e Y. El ajuste es posible porque tanto $n_x P_x Q_x$ como $n_y P_y Q_y$ son mayores que 3.

De este modo, la probabilidad que se pide puede expresarse como:

$$r_x(HPGMT) \rightarrow X(Normal) \qquad r_y(HPGMT) \rightarrow Y(Normal)$$

$$E[X] = E[r_x] = 130 \qquad E[Y] = E[r_y] = 130$$

$$Var(x) = Var(r_x) = 45.5 \qquad Var(y) = Var(r_y) = 45.5$$

$$X \approx N(130;45.5) \qquad Y \approx N(130;45.5)$$

$$P[|r_x - r_y| > 10] = P[r_x - r_y > 10] + P[r_y - r_x > 10]$$

$$= P[X-Y > 10.5] + P[Y-X > 10.5]$$

Tanto X-Y como Y-X son combinaciones lineales de dos variables normales cuya naturaleza y parámetros son:

$$Nat(X-Y) \rightarrow NORMAL \leftarrow Nat(Y-X)$$

$$E[X-Y] = E[Y-X] = 130-130 = 0$$

$$Var(X-Y) = Var(Y-X) = VarX + VarY =$$

$$= 45.5+45.5 = 91$$

Por tanto, la probabilidad pedida será:

$$P[X-Y > 10.5] + P[Y-X > 10.5] =$$

$$= P[Z > \frac{10.5-0}{\sqrt{91}}] + P[Z > \frac{10.5-0}{\sqrt{91}}]$$

$$= 2\ P[Z > 1.1] = 0.2714$$